Preface
Introductory Statistics is intended for the one-semester introduction to statistics course for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real world settings, and assumes that students have an understanding of intermediate algebra. In addition to end of section practice and homework sets, examples of each topic are explained step-by-step throughout the text and followed by a Try It problem that is designed as extra practice for students. This book also includes collaborative exercises and statistics labs designed to give students the opportunity to work together and explore key concepts. To support today's student in understanding technology, this book features TI 83, 83+, 84, or 84+ calculator instructions at strategic points throughout. While the book has been built so that each chapter builds on the previous, it can be rearranged to accommodate any instructor's particular needs.

Welcome to *Introductory Statistics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

The foundation of this textbook is *Collaborative Statistics*, by Barbara Illowsky and Susan Dean. Additional topics, examples, and innovations in terminology and practical applications have been added, all with a goal of increasing relevance and accessibility for students.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 25 books for college and AP® courses used by hundreds of thousands of students. OpenStax Tutor, our low-cost personalized learning tool, is being used in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax's resources

Customization

Introductory Statistics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit your book page on OpenStax.org for more information.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on OpenStax.org.

Format

You can access this textbook for free in web view or PDF through OpenStax.org, and in low-cost print and iBooks editions.

About *Introductory Statistics*

Introductory Statistics follows scope and sequence requirements of a one-semester introduction to statistics course and is geared toward students majoring in fields other than math or engineering. The text assumes some knowledge of intermediate algebra and focuses on statistics application over theory. *Introductory Statistics* includes innovative practical applications that make the text relevant and accessible, as well as collaborative exercises, technology integration problems, and statistics labs.

Coverage and scope

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 Continuous Random Variables
Chapter 6 The Normal Distribution
Chapter 7 The Central Limit Theorem
Chapter 8 Confidence Intervals
Chapter 9 Hypothesis Testing with One Sample
Chapter 10 Hypothesis Testing with Two Samples
Chapter 11 The Chi-Square Distribution
Chapter 12 Linear Regression and Correlation
Chapter 13 F Distribution and One-Way ANOVA

Alternate sequencing

Introductory Statistics was conceived and written to fit a particular topical sequence, but it can be used flexibly to accommodate other course structures. One such potential structure, which fits reasonably well with the

textbook content, is provided below. Please consider, however, that the chapters were not written to be completely independent, and that the proposed alternate sequence should be carefully considered for student preparation and textual consistency.

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 12 Linear Regression and Correlation
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 Continuous Random Variables
Chapter 6 The Normal Distribution
Chapter 7 The Central Limit Theorem
Chapter 8 Confidence Intervals
Chapter 9 Hypothesis Testing with One Sample
Chapter 10 Hypothesis Testing with Two Samples
Chapter 11 The Chi-Square Distribution
Chapter 13 F Distribution and One-Way ANOVA

Pedagogical foundation and features

- **Examples** are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics, including examples about college life and learning, health and medicine, retail and business, and sports and entertainment.
- **Try It** practice problems immediately follow many examples and give students the opportunity to practice as they read the text. **They are usually based on practical and familiar topics, like the Examples themselves**.
- **Collaborative Exercises** provide an in-class scenario for students to work together to explore presented concepts.
- **Using the TI-83, 83+, 84, 84+ Calculator** shows students step-by-step instructions to input problems into their calculator.

The Technology Icon indicates where the use of a TI calculator or computer software is recommended.
Practice, Homework, and Bringing It Together problems give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Statistics labs

These innovative activities were developed by Barbara Illowsky and Susan Dean in order to offer students the experience of designing, implementing, and interpreting statistical analyses. They are drawn from actual experiments and data-gathering processes and offer a unique hands-on and collaborative experience. The labs provide a foundation for further learning and classroom interaction that will produce a meaningful application of statistics.

Statistics Labs appear at the end of each chapter and begin with student learning outcomes, general estimates for time on task, and any global implementation notes. Students are then provided with step-by-step guidance, including sample data tables and calculation prompts. The detailed assistance will help the students successfully apply the concepts in the text and lay the groundwork for future collaborative or individual work.

Additional resources

Student and instructor resources

We’ve compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Community Hubs

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons – a platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty.

To reach the Community Hubs, visit www.oercommons.org/hubs/OpenStax.

Partner resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on OpenStax.org.

About the authors

Senior contributing authors

Barbara Illowsky, De Anza College
Susan Dean, De Anza College

Contributing authors

Birgit Aquilonius, West Valley College
Charles Ashbacher, Upper Iowa University, Cedar Rapids

Abraham Biggs, Broward Community College
Daniel Birmajer, Nazareth College
Roberta Bloom, De Anza College
Bryan Blount, Kentucky Wesleyan College
Ernest Bonat, Portland Community College
Sarah Boslaugh, Kennesaw State University
David Bosworth, Hutchinson Community College
Sheri Boyd, Rollins College
George Bratton, University of Central Arkansas
Jing Chang, College of Saint Mary
Laurel Chiappetta, University of Pittsburgh
Lenore Desilets, De Anza College
Matthew Einsohn, Prescott College
Ann Flanigan, Kapiolani Community College
David French, Tidewater Community College
Mo Geraghty, De Anza College
Larry Green, Lake Tahoe Community College
Michael Greenwich, College of Southern Nevada
Inna Grushko, De Anza College
Valier Hauber, De Anza College
Janice Hector, De Anza College
Jim Helmreich, Marist College
Robert Henderson, Stephen F. Austin State University
Mel Jacobsen, Snow College
Mary Jo Kane, De Anza College
Lynette Kenyon, Collin County Community College
Charles Klein, De Anza College
Alexander Kolovos
Sheldon Lee, Viterbo University
Sara Lenhart, Christopher Newport University
Wendy Lightheart, Lane Community College
Vladimir Logvenenko, De Anza College
Jim Lucas, De Anza College
Lisa Markus, De Anza College
Miriam Masullo, SUNY Purchase
Diane Mathios, De Anza College
Robert McDevitt, Germanna Community College

Mark Mills, Central College
Cindy Moss, Skyline College
Nydia Nelson, St. Petersburg College
Benjamin Ngwudike, Jackson State University
Jonathan Oaks, Macomb Community College
Carol Olmstead, De Anza College
Adam Pennell, Greensboro College
Kathy Plum, De Anza College
Lisa Rosenberg, Elon University
Sudipta Roy, Kankakee Community College
Javier Rueda, De Anza College
Yvonne Sandoval, Pima Community College
Rupinder Sekhon, De Anza College
Travis Short, St. Petersburg College
Frank Snow, De Anza College
Abdulhamid Sukar, Cameron University
Jeffery Taub, Maine Maritime Academy
Mary Teegarden, San Diego Mesa College
John Thomas, College of Lake County
Philip J. Verrecchia, York College of Pennsylvania
Dennis Walsh, Middle Tennessee State University
Cheryl Wartman, University of Prince Edward Island
Carol Weideman, St. Petersburg College
Andrew Wiesner, Pennsylvania State University

Introduction
class="introduction"

We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

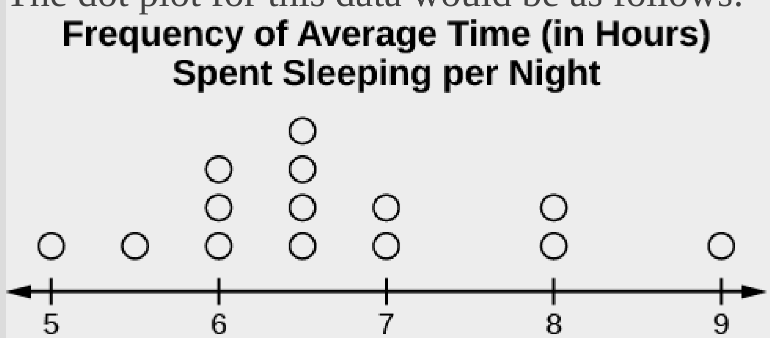
You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

Note:
Collaborative Exercise
In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:
5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9
The dot plot for this data would be as follows:


Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not? Where do your data appear to cluster? How might you interpret the clustering?
The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example,

finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or

not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in

inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, usually notated by capital letters such as X and Y, is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:
NOTE
The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Note:
Try It

Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Solution:
Try It Solutions

The **population** is all families with children attending Knoll Academy.

The **sample** is a random selection of 100 families with children attending Knoll Academy.

The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.

The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.

The **data** are the dollar amounts spent by the families. Examples of the data are \$65, \$75, and \$95.

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA’s of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population_____ 2. Statistic _____ 3. Parameter _____ 4. Sample _____ 5. Variable _____ 6. Data _____

- a) all students who attended the college last year
- b) the cumulative GPA of one student who graduated from the college last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the college last year, randomly selected
- e) the average cumulative GPA of students who graduated from the college last year
- f) all students who graduated from the college last year
- g) the average cumulative GPA of students in the study who graduated from the college last year

Solution:

1. f 2. g 3. e 4. d 5. b 6. c

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution:

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution:

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Note:
Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

Chapter Review

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:
3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B:
3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Determine what the key terms refer to in the example for Researcher A.

Exercise:

Problem: population

Solution:

AIDS patients.

Exercise:

Problem: sample

Exercise:

Problem: parameter

Solution:

The average length of time (in months) AIDS patients live after treatment.

Exercise:

Problem: statistic

Exercise:

Problem: variable

Solution:

X = the length of time (in months) AIDS patients live after treatment

HOMEWORK

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

Exercise:

Problem:

A fitness center is interested in the mean amount of time a client exercises in the center each week.

Exercise:

Problem:

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

Solution:

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

Exercise:

Problem:

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Exercise:

Problem:

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

Solution:

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

Exercise:

Problem:

A politician is interested in the proportion of voters in his district who think he is doing a good job.

Exercise:

Problem:

A marriage counselor is interested in the proportion of clients she counsels who stay married.

Solution:

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor’s clients who stay married
e. X = the number of couples who stay married
f. yes, no

Exercise:

Problem:

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Exercise:

Problem:

A marketing company is interested in the proportion of people who will buy a particular product.

Solution:

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

Exercise:

Problem: Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

Solution:

a

Exercise:

Problem:

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

Glossary

Average
: also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic
a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable
a characteristic of interest for each person or object in a population

Frequency, Frequency Tables, and Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative(categorical)**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces

of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 56332475235654435253.

[link] lists the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6

DATA VALUE	FREQUENCY
6	2
7	1

Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to [link], there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of [link] is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in [link].

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note:
NOTE
Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

[link] represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100}$ = 0.05	0.05
61.95–63.95	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
63.95–65.95	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
65.95–67.95	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63
67.95–69.95	17	$\frac{17}{100}$ = 0.17	0.63 + 0.17 = 0.80
69.95–71.95	12	$\frac{12}{100}$ = 0.12	0.80 + 0.12 = 0.92
71.95–73.95	7	$\frac{7}{100}$ = 0.07	0.92 + 0.07 = 0.99
73.95–75.95	1	$\frac{1}{100}$ = 0.01	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.95 inches
- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

Note:
Note
This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example:
Exercise:

Problem: From [link], find the percentage of heights that are less than 65.95 inches.

Solution:

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Note:
Try It
Exercise:

Problem: [link] shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50}$ = 0.12	0.12
4.97–6.99	7	$\frac{7}{50}$ = 0.14	0.12 + 0.14 = 0.26
6.99–9.01	15	$\frac{15}{50}$ = 0.30	0.26 + 0.30 = 0.56
9.01–11.03	8	$\frac{8}{50}$ = 0.16	0.56 + 0.16 = 0.72
11.03–13.05	9	$\frac{9}{50}$ = 0.18	0.72 + 0.18 = 0.90
13.05–15.07	5	$\frac{5}{50}$ = 0.10	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

From [link], find the percentage of rainfall that is less than 9.01 inches.

Solution:
Try It Solutions

0.56 or 56%

Example:
Exercise:

Problem:

From [link], find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution:

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Note:
Try It
Exercise:

Problem: From [link], find the percentage of rainfall that is between 6.99 and 13.05 inches.

Solution:
Try It Solutions

0.30 + 0.16 + 0.18 = 0.64 or 64%

Example:
Exercise:

Problem:

Use the heights of the 100 male semiprofessional soccer players in [link]. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the

previous relative frequencies to the relative frequency for the current row.

Solution:

a. 29%
b. 36%
c. 77%
d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

Note:
Try It
Exercise:

Problem:

From [link], find the number of towns that have rainfall between 2.95 and 9.01 inches.

Solution:
Try It Solutions

6 + 7 + 15 = 28 towns

Note:
Collaborative Exercise
In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class have no siblings?
2. What percentage of the students have from one to three siblings?
3. What percentage of the students have fewer than three siblings?

Example:
Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2 5 7 3 2 10 18 15 20 7 10 18 5 12 13 12 4 5 10. [link] was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Frequency of Commuting Distances

Exercise:

Problem:

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution:

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
c. $\frac{5}{19}$

d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Note:
Try It
Exercise:

Problem:

[link] represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Solution:
Try It Solutions

$\frac{9}{50}$

Example:
[link] contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712

Year	Total Number of Deaths
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Exercise:

Problem: Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution:

a. 97,118 (11.8%)
b. 41.6%
c. 67,092/823,356 or 0.081 or 8.1 %
d. 27.8%
e. Quantitative discrete
f. Quantitative continuous

Note:
Try It
Exercise:

Problem:

[link] contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Solution:
Try It Solutions

a. 190,800 (29.2%)
b. 24.9%
c. 260,086/653,782 or 39.8%
d. 4.6%
e. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

"State & County QuickFacts," U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

"State & County QuickFacts: Quick, easy access to facts about people, business, and geography," U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).

"Table 5: Direct hits by mainland United States Hurricanes (1851-2004)," National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).

"Levels of Measurement," http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, "Levels of Measurement," about.com, http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm (accessed May 1, 2013).

David Lane. "Levels of Measurement," Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Chapter Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

Exercise:

Problem: What type of measure scale is being used? Nominal, ordinal, interval or ratio.

a. High school soccer players classified by their athletic ability: Superior, Average, Above average
b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
c. The colors of crayons in a 24-crayon box

d. Social security numbers
e. Incomes measured in dollars
f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
g. Political outlook: extreme left, left-of-center, right-of-center, extreme right
h. Time of day on an analog watch
i. The distance in miles to the closest grocery store
j. The dates 1066, 1492, 1644, 1947, and 1944
k. The heights of 21–65 year-old women
l. Common letter grades: A, B, C, D, and F

Solution:

a. ordinal
b. interval
c. nominal
d. nominal
e. ratio
f. ordinal
g. nominal
h. interval
i. ratio
j. interval
k. ratio
l. ordinal

HOMEWORK

Exercise:

Problem:

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
3			

Part-time Student Course Loads

a. Fill in the blanks in [link].
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

Exercise:

Problem:

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in [link].

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in [link].
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

Solution:

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

Exercise:

Problem:

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2 5 7 2 2 10 20 15 0 7 0 20 5 12 15 12 4 5 10 .

[link] was produced.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Frequency of Immigrant Survey Responses

a. Fix the errors in [link]. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: "47 percent of the people surveyed have lived in the U.S. for 5 years."
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

Exercise:

Problem:

How much time does it take to travel to work? [link] shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

Solution:

The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

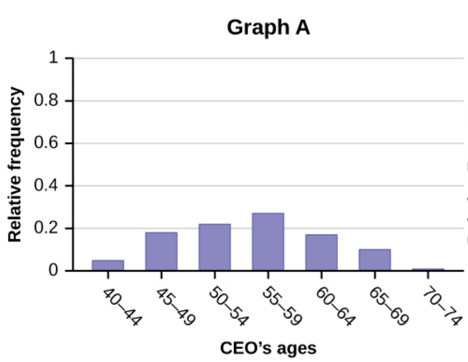
Exercise:

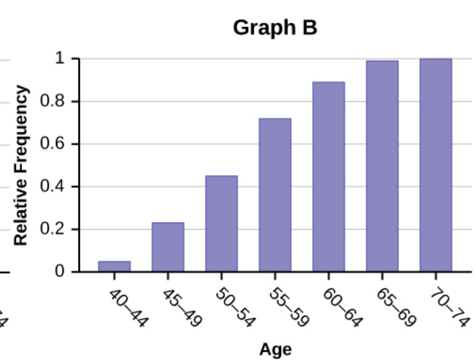
Problem:

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. [link] shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?





Use the following information to answer the next two exercises: [link] contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Frequency of Hurricane Direct Hits

Exercise:

Problem: What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

Solution:

b

Exercise:

Problem:

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

Glossary

Cumulative Relative Frequency
    The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Frequency
    the number of times a value of the data occurs

Relative Frequency
    the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Introduction
class="introduction"

When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

Note:
NOTE
This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The [Texas Instruments (TI) website]() provides additional instructions for using these calculators.

Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:
Equation:

$$\mathrm{RF} = \frac{f}{n}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and RF = $\frac{f}{n}$ = $\frac{3}{40}$ = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example:
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.
60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5
66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.
60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.
The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.
Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

Equation:

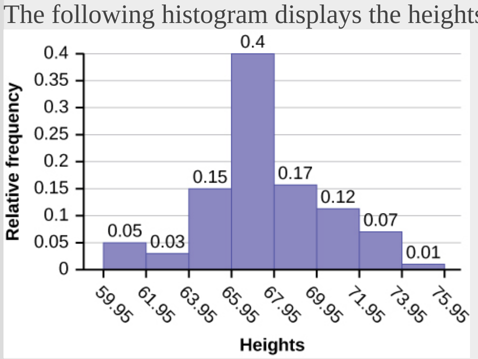
$$\frac{74.05 - 59.95}{8} = 1.76$$

Note:
NOTE
We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the number of bars or class intervals is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.
The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.



Note:
Try It
Exercise:

Problem:

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.
9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Solution:

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

$\frac{14.05-8.95}{6} = 0.85$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Example:
Create a histogram for the following data: the number of books bought by 50 part-time college students at ABC College.the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.
1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6
Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.
Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.
Exercise:

Problem:

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _______ to _______, the 5 in the middle of the interval from _______ to _______, and the _______ in the middle of the interval from _______ to _______ .

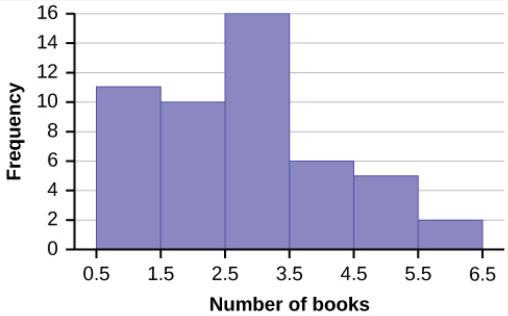
Solution:

- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:
Equation:

$$\frac{6.5 - 0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.
The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.



Note:
Go to [link]. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for [link].

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xmax = 6.5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2$^{\text{nd}}$ Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2$^{\text{nd}}$ Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3$^{\text{rd}}$ picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2$^{\text{nd}}$ 1). Arrow down to Freq. Enter L2 (2$^{\text{nd}}$ 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

Note:
Try It
Exercise:

Problem:

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3
20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

Solution:

1.5
1.5 to 2.5
2.5 to 3.5

Example:
Exercise:

Problem: Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Solution:

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Note:
Try It
Exercise:

Problem:

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22351526 40281820 25343942 24221927 22344020 38and 28
Use 10–19 as the first interval.

Note:
Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

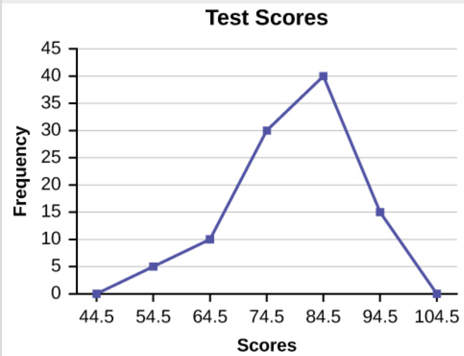
To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example:
A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
79.5	89.5	40	85
89.5	99.5	15	100



The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

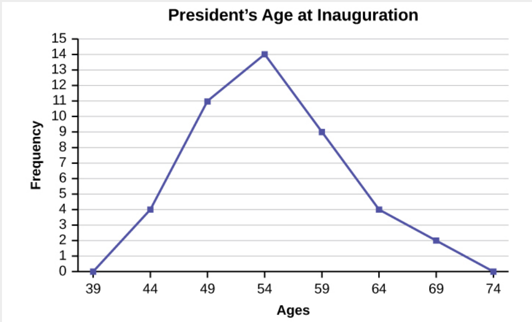
Note:
Try It
Exercise:

Problem: Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in [link].

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Solution:

The first label on the *x*-axis is 39. This represents an interval extending from 36.5 to 41.5. Since there are no ages less than 41.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 44 represents the next interval, or the first “real” interval from the table, and contains four scores. This reasoning is followed for each of the remaining intervals with the point 74 representing the interval from 71.5 to 76.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

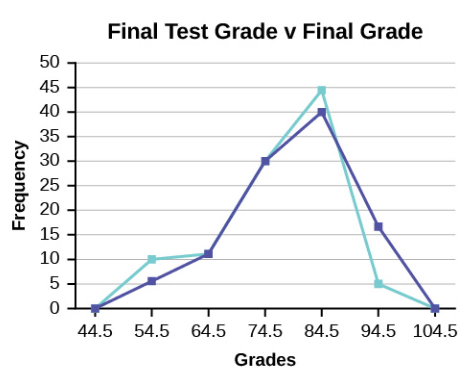


Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example:
We will construct an overlay frequency polygon comparing the scores from [link] with the students’ final numeric grade.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100



Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don‘t have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example:
Exercise:

Problem:

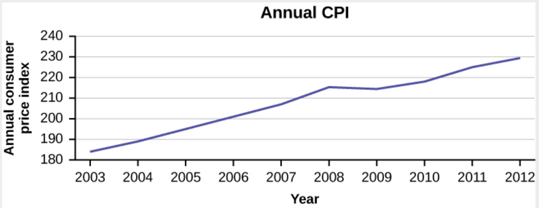
The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056

Year	Aug	Sep	Oct	Nov	Dec	Annual
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Solution:



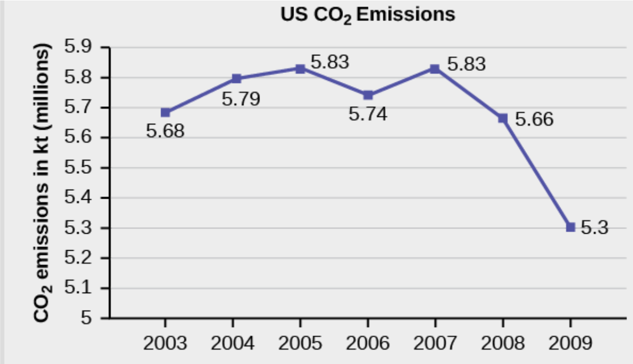
Note:
Try It
Exercise:

Problem:

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO2 Emissions			
	Ukraine	**United Kingdom**	**United States**
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Solution:



Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

References

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Chapter Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

Exercise:

Problem:

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Exercise:

Problem: What does the frequency column in [link] sum to? Why?

Solution:

65

Exercise:

Problem: What does the relative frequency column in [link] sum to? Why?

Exercise:

Problem: What is the difference between relative frequency and frequency for each data value in [link]?

Solution:

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

Exercise:

Problem:

What is the difference between cumulative relative frequency and relative frequency for each data value?

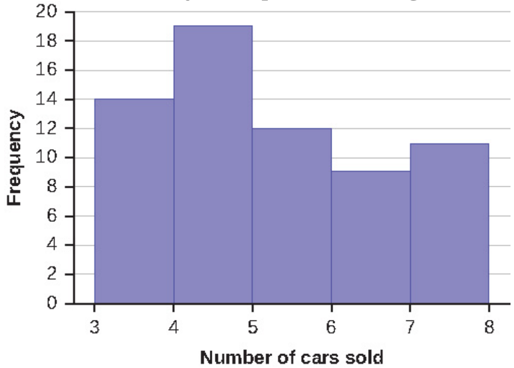
Exercise:

Problem:

To construct the histogram for the data in [link], determine appropriate minimum and maximum *x* and *y* values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Solution:

Answers will vary. One possible histogram is shown:



Exercise:

Problem: Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Exercise:

Problem:

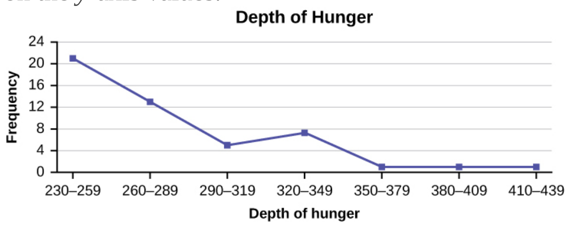
Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1

Depth of Hunger	Frequency
410–439	1

Solution:

Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.



Exercise:

Problem:

Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3

Life Expectancy at Birth – Men	**Frequency**
63–69	1
70–76	1
77–83	7
84–90	5

Exercise:

Problem:

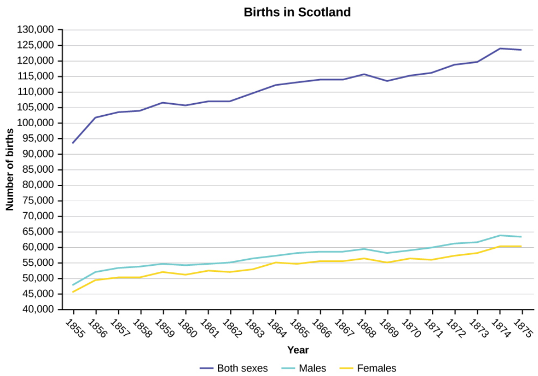
Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	18
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	11

Sex/Year	1870	1871	1872	1873	1874	1875
Female	56,431	56,099	57,472	58,233	60,109	60,146
Male	58,959	60,029	61,293	61,467	63,602	63,432
Total	115,390	116,128	118,765	119,700	123,711	123,578

Solution:



Exercise:

Problem:

The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

a. Construct a double time series graph using a common x-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

Homework

Exercise:

Problem:

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher B

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise:

Problem:

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Amount($)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Singles

Amount($)	Frequency	Rel. Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
c. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
d. Compare the two graphs:
   i. List two similarities between the graphs.
   ii. List two differences between the graphs.
   iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
f. Compare the graph for the singles with the new graph for the couples:
   i. List two similarities between the graphs.
   ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

Solution:

Amount($)	Frequency	Relative Frequency

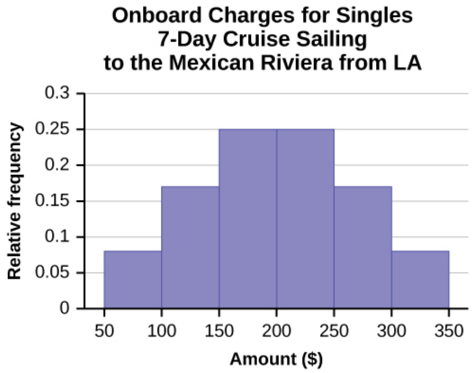
Amount($)	Frequency	Relative Frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Singles

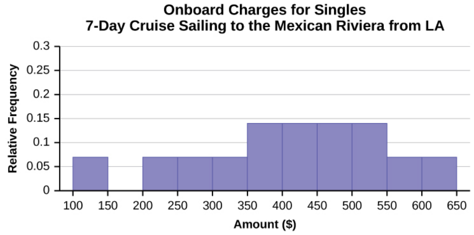
Amount($)	Frequency	Relative Frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Couples

a. See [link] and [link].
b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).



c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).



d. Compare the two graphs:

   i. Answers may vary. Possible answers include:

      - Both graphs have a single peak.
      - Both graphs use class intervals with width equal to $50.

   ii. Answers may vary. Possible answers include:

      - The couples graph has a class interval with no values.
      - It takes almost twice as many class intervals to display the data for couples.

   iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.
f. Compare the graph for the Singles with the new graph for the Couples:

   i.
      - Both graphs have a single peak.
      - Both graphs display 6 class intervals.
      - Both graphs show the same general pattern.

   ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.
h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

Exercise:

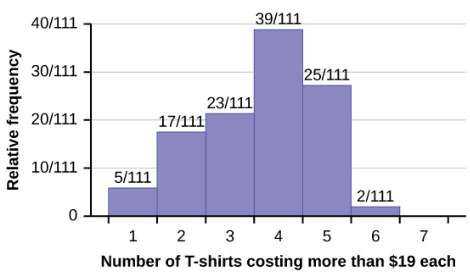
Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.



Exercise:

Problem:

The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21
b. 59
c. 41
d. Cannot be determined

Solution:

c

Exercise:

Problem:

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

Exercise:

Problem: Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5

State	Percent (%)	State	Percent (%)	State	Percent (%)
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the *x*-axis with the states.

Solution:

Answers will vary.

Glossary

Frequency
: the number of times a value of the data occurs

Histogram
: a graphical representation in *x*-*y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency
: the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note:
NOTE
The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an *x* with a bar over it (pronounced “*x* bar”): x.

The Greek letter *μ* (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:
1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4
Equation:

$$x=\frac{1+1+1+2+2+3+4+4+4+4+4}{11}=2.7$$

Equation:

$$x=\frac{3(1)+2(2)+1(3)+5(4)}{11}=2.7$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter *n* is the total number of data values in the sample. If *n* is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If *n* is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2}=\frac{97+1}{2}=49$. The median is the 49^{th} value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2}=\frac{100+1}{2}=50.5$. The median occurs midway between the 50^{th} and 51^{st} values. The location of the median and the value of the median are **not** the same. The upper case letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:
Exercise:

Problem:

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;
Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$x = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, *M*, first use the formula for the location. The location is:
$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$
Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24+24}{2} = 24$$

Note:
To find the mean and the median:
Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.
Enter data into the list editor. Press STAT 1:EDIT.
Put the data values into list L1.
Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
Press the down and up arrow keys to scroll.
x = 23.6, M = 24

Note:
Try It
Exercise:

Problem:

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3 4 5 7 7 7 7 8 8 9 9 10 10 10 10 10 11 12 12 13 14 14 15 15 17 17 18 19 19 19 21 21 22 22 23 24 24 24 24

Solution:

Mean: 3 + 4 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 10 + 10 + 10 + 10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 15 + 15 + 17 + 17 + 18 + 19 + 19 + 19 + 21 + 21 + 22 + 22 + 23 + 24 + 24 + 24 = 544
$\frac{544}{39} = 13.95$
Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example:
Exercise:

Problem:

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution:

$x = \frac{5{,}000{,}000+49(30{,}000)}{50} = 129{,}400$

M = 30,000

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Note:
Try It
Exercise:

Problem:

In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the “center”: the mean or the median?

Solution:

The median is the better measure of the “center” than the mean because 59 of the values are $280,000 and one is $2,500,000. The $2,500,000 is an outlier. Either $280,000 or $315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example:
Statistics exam scores for 20 students are as follows:
50535959636372727272727678818384848490 93
Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Note:
Try It
Exercise:

Problem: The number of books checked out from the library from 25 students are as follows:

0001233445577778889101011111212
Find the mode.

Solution:

The most frequent number of books is 7, which occurs four times. Mode = 7.

Example:
Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.
When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note:
NOTE
The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Note:
Try It

Exercise:

Problem:

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the “center”?

Solution:

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don’t represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean x of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$

# of movies	Relative Frequency
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean x is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: *mean* = $\frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary+upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where *f* = the frequency of the interval and *m* = the midpoint of the interval.

Example:
Exercise:

Problem:

A frequency table displaying professor Blount’s last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution:

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

  $53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$

- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Note:
Try It
Exercise:

Problem:

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Solution:

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers
The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.
Mean $= \frac{(1.75)(3)+(5.5)(7)+(9.5)(12)+(13.5)(7)+(17.5)(9)}{(3+7+12+7+9)} = \frac{409.75}{38} = 10.78$

References

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Chapter Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$\mu = \frac{\sum fm}{\sum f}$ Where *f* = interval frequencies and *m* = interval midpoints.

Exercise:

Problem: Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1

Daily Low Temperature	Frequency
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 161719202021232425252526262727272829303233333435373940

Exercise:

Problem: Calculate the mean.

Solution:

Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738;

$\frac{738}{27}$ = 27.33

Exercise:

Problem: Identify the median.

Exercise:

Problem: Identify the mode.

Solution:

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car

salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:
Exercise:

Problem: sample mean = $\bar{x}$ = _______

Exercise:

Problem: median = _______

Solution:

4

Exercise:

Problem: mode = _______

Homework

Exercise:

Problem:

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

Exercise:

Problem:

[link] gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	**Number of Countries**
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Solution:

The mean percentage, $x = \frac{1328.65}{50} = 26.75$

Bringing It Together

Exercise:

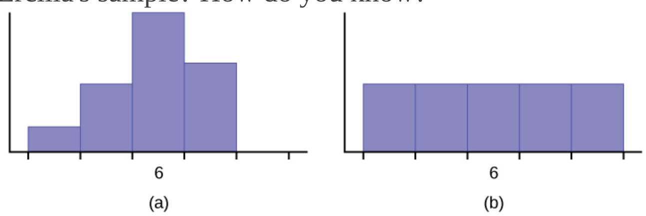
Problem:

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

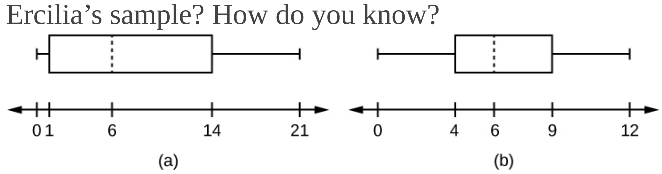
	Javier	**Ercilia**

	Javier	Ercilia
x	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?



Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Exercise:

Problem: What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Solution:

a

Exercise:

Problem: What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise:

Problem: Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Solution:

b

Glossary

Frequency Table
: a data representation in which grouped data is displayed along with the corresponding frequencies

Mean
: a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median
: a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may

or may not be part of the data.

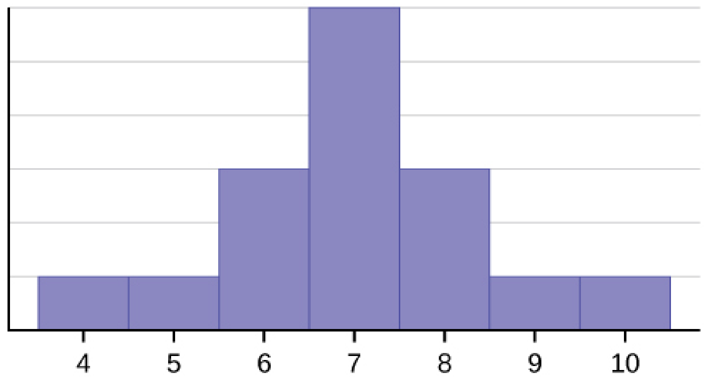
Midpoint
: the mean of an interval in a frequency table

Mode
: the value that appears most frequently in a set of data

Skewness and the Mean, Median, and Mode

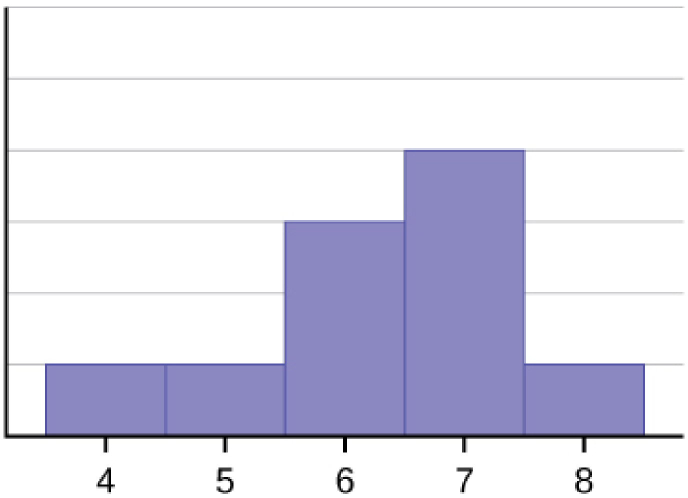
Consider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.



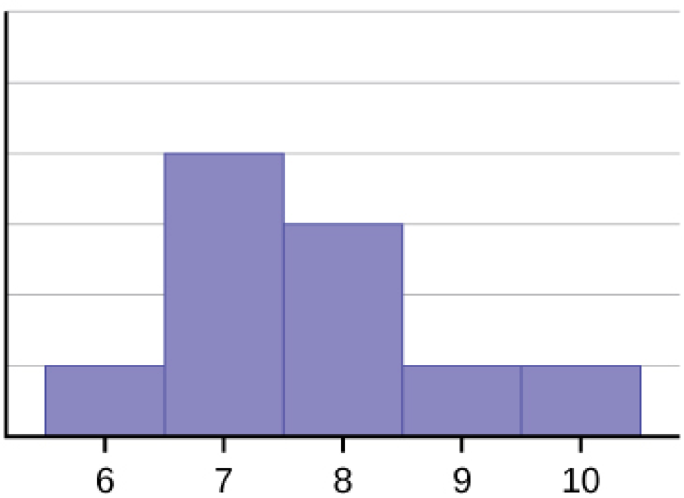
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4566677778 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.



The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 67777888910, is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the

distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example:
Exercise:

Problem:

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

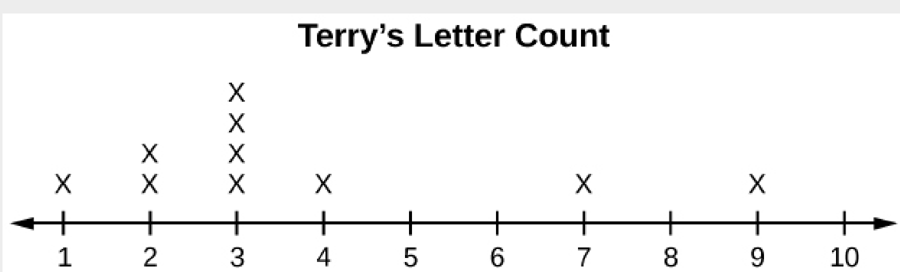
Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2

Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1

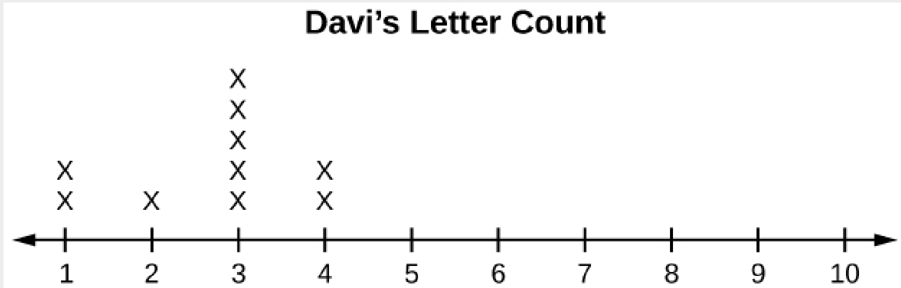
Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

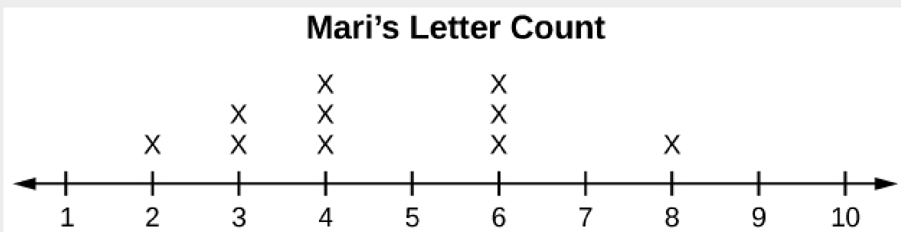
Solution:

a. 

Terry’s distribution has a right (positive) skew.



Davis’ distribution has a left (negative) skew



Maris’ distribution is symmetrically shaped.

b. Terry’s mean is 3.7, Davis’ mean is 2.7, Maris’ mean is 4.6.
c. Terry’s median is three, Davis’ median is three. Maris’ median is four.
d. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Note:
Try It
Exercise:

Problem:

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

c.

Solution:

a. mean = 4.25, median = 3.5, mode = 1; The mean > median > mode which indicates skewness to the right. (data are 0, 1, 2, 3, 4, 5, 6, 9, 10, 14 and respective frequencies are 2, 4, 3, 1, 2, 2, 2, 2, 1, 1)
b. mean = 70.1 , median = 68, mode = 57, 67 bimodal; the mean and median are close but there is a little skewness to the right which is influenced by the data being bimodal. (data are 46, 49, 53, 56, 57, 57, 57, 58, 60, 60, 63, 63, 64, 64, 65, 66, 67, 67, 67, 68, 70, 71, 71, 72, 73, 74, 77, 78, 78, 79, 80, 81, 83, 85, 88, 90, 90 93, 93).
c. These are estimates: mean =16.095, median = 17.495, mode = 22.495 (there may be no mode); The mean < median < mode which indicates skewness to the left. (data are the midponts of the intervals: 2.495, 7.495, 12.495, 17.495, 22.495 and respective frequencies are 2, 3, 4, 7, 9).

Chapter Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of

distributions. A **right (or positive) skewed** distribution has a shape like [link]. A **left (or negative) skewed** distribution has a shape like [link]. A **symmetrical** distrubtion looks like [link].

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

Exercise:

Problem: 1112222333333344455

Solution:

The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

Exercise:

Problem:16171922222222223

Exercise:

Problem:87878787878889899091

Solution:

The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercise:

Problem:

When the data are skewed left, what is the typical relationship between the mean and median?

Exercise:

Problem:

When the data are symmetrical, what is the typical relationship between the mean and median?

Solution:

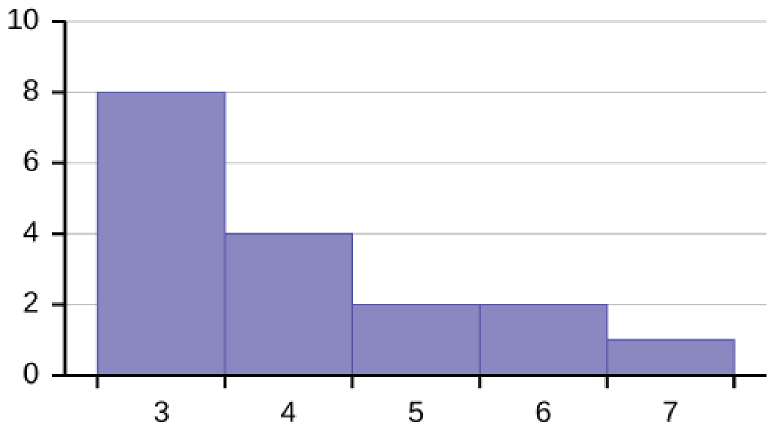
When the data are symmetrical, the mean and median are close or the same.

Exercise:

Problem: What word describes a distribution that has two modes?

Exercise:

Problem: Describe the shape of this distribution.



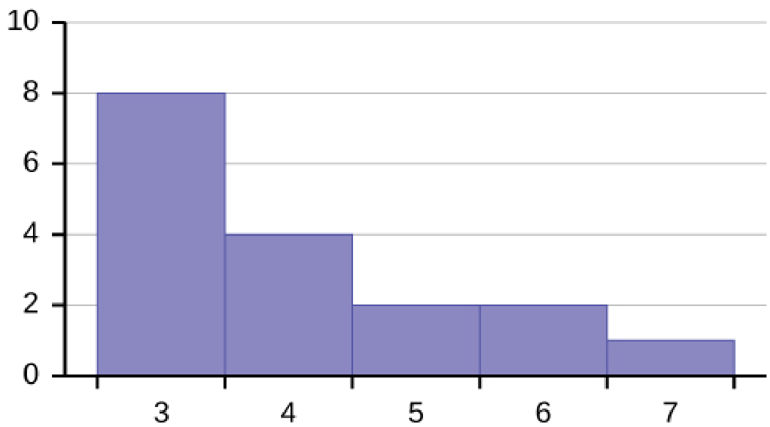
Solution:

The distribution is skewed right because it looks pulled out to the right.

Exercise:

Problem:

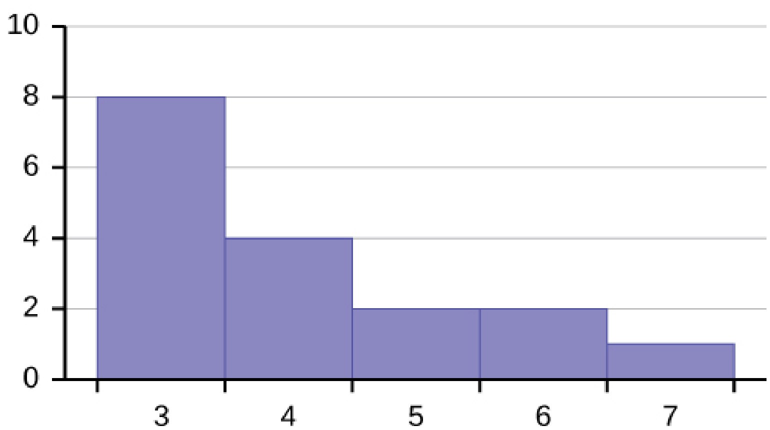
Describe the relationship between the mode and the median of this distribution.



Exercise:

Problem:

Describe the relationship between the mean and the median of this distribution.

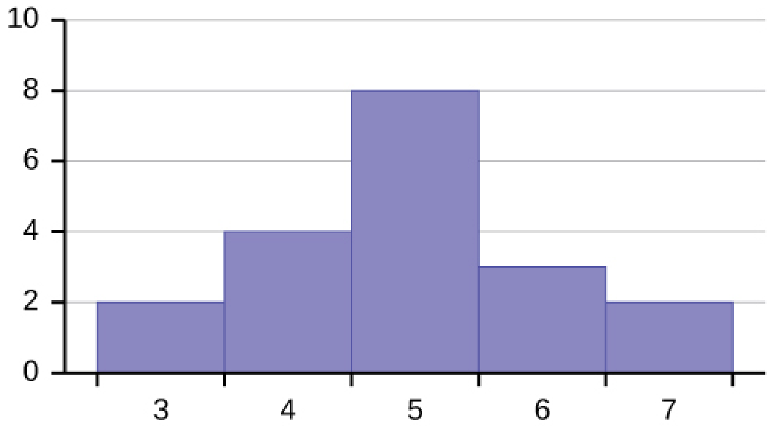


Solution:

The mean is 4.1 and is slightly greater than the median, which is four.

Exercise:

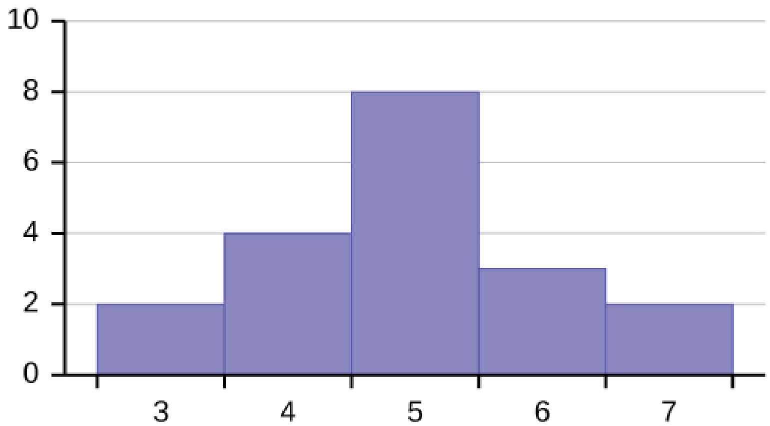
Problem: Describe the shape of this distribution.



Exercise:

Problem:

Describe the relationship between the mode and the median of this distribution.



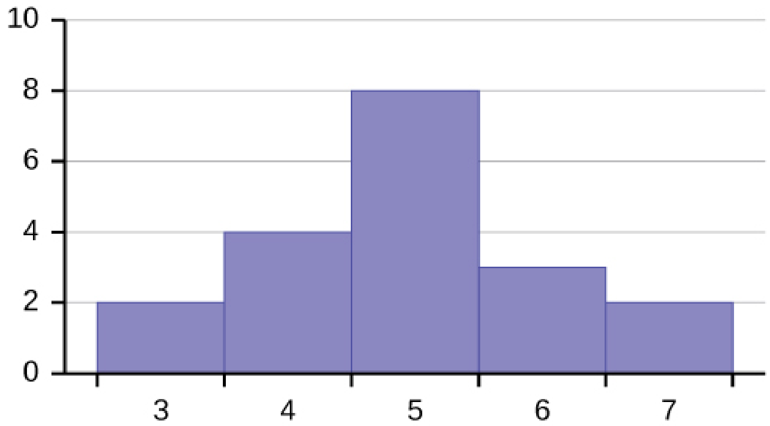
Solution:

The mode and the median are the same. In this case, they are both five.

Exercise:

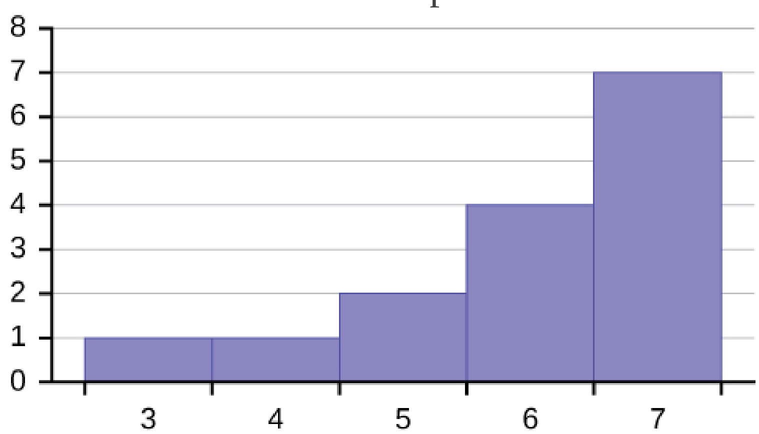
Problem:

Are the mean and the median the exact same in this distribution? Why or why not?



Exercise:

Problem: Describe the shape of this distribution.



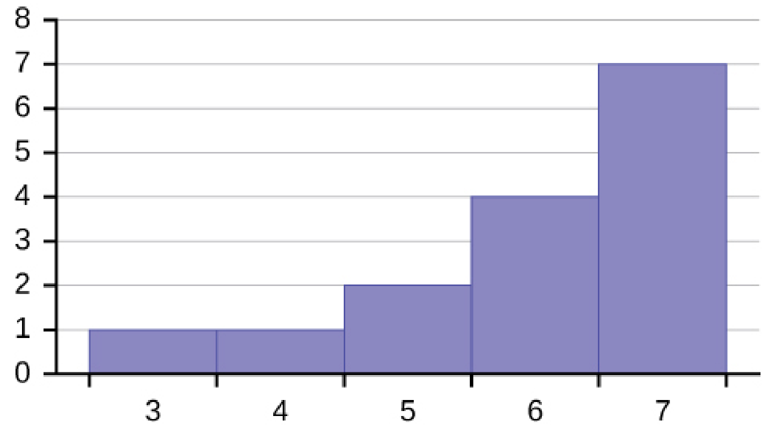
Solution:

The distribution is skewed left because it looks pulled out to the left.

Exercise:

Problem:

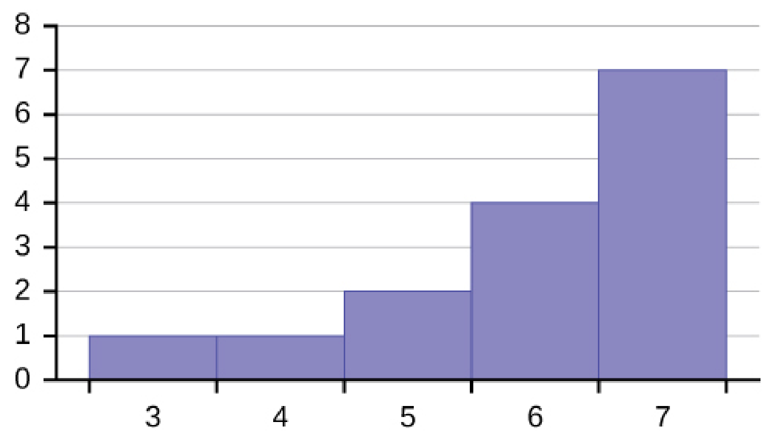
Describe the relationship between the mode and the median of this distribution.



Exercise:

Problem:

Describe the relationship between the mean and the median of this distribution.



Solution:

The mean and the median are both six.

Exercise:

Problem: The mean and median for the data are the same.

345566667777777

Is the data perfectly symmetrical? Why or why not?

Exercise:

Problem:

Which is the greatest, the mean, the mode, or the median of the data set?

11111212121212131517222222

Solution:

The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.

Exercise:

Problem:

Which is the least, the mean, the mode, and the median of the data set?

56565658596062646467

Exercise:

Problem:

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Solution:

The mean tends to reflect skewing the most because it is affected the most by outliers.

Exercise:

Problem:

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

Homework

Exercise:

Problem:

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because 5 + (1)(2) = 7.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because 5 + (–2)(2) = 1.

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: 1 = 5 + (–2)(2).

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population.

- **sample:** $x = x + (\#ofSTDEV)(s)$
- **Population:** $x = \mu + (\#ofSTDEV)(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol x is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - x$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - x$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by ***n* – 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n - 1}}$
- For the sample standard deviation, the denominator is ***n* - 1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}}$
- For the population standard deviation, the denominator is *N*, the number of items in the population.

In these formulas, *f* represents the frequency with which a value appears. For example, if a value appears once, *f* is one. If a value appears three times in the data set or population, *f* is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Descriptive Statistics: Measuring the Center of the Data. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in the chapter The Central Limit Theorem (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note:
NOTE
In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example:

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:
9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

Equation:

$$x = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.
The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating *s*.

Data	**Freq.**	**Deviations**	***Deviations*^{2}**	**(Freq.) (*Deviations*^{2})**
x	f	$(x - x)$	$(x - x)^2$	$(f)(x - x)^2$
9	1	9 – 10.525 = –1.525	$(-1.525)^2$ = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(-1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(-0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(-0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875

Data	Freq.	Deviations	*Deviations*^{2}	(Freq.)(*Deviations*^{2})
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$s^2 = \frac{9.7375}{20-1} = 0.5125$

The **sample standard deviation** s is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, s = 0.72.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Exercise:

Problem:

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: x = $\bar{x}$ + (#ofSTDEVs)(s)
- For a population: x = μ + (#ofSTDEVs)(σ)
- For this example, use x = $\bar{x}$ + (#ofSTDEVs)(s) because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
c. Find the value that is two standard deviations below the mean. Find ($\bar{x}$ – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

a. **Note:**

   - Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
   - Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.

- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- x = 10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

b. (x + 1s) = 10.53 + (1)(0.72) = 11.25
c. (x – 2s) = 10.53 – (2)(0.72) = 9.09

d.
- (x – 1.5s) = 10.53 – (1.5)(0.72) = 9.45
- (x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61

Note:
Try It
Exercise:

Problem: On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Solution:

μ = 30.68

s = 6.09
(x + 2s) = 30.68 + (2)(6.09) = 42.86.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For [link], there are n = 20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n = 20, the calculation divided by n – 1 = 20 – 1 = 19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n – 1). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n – 1) gives a better estimate of the population variance.

Note:
NOTE
Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that

the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example:
Exercise:

Problem:

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:

    i. The sample mean
    ii. The sample standard deviation
    iii. The median
    iv. The first quartile
    v. The third quartile
    vi. *IQR*

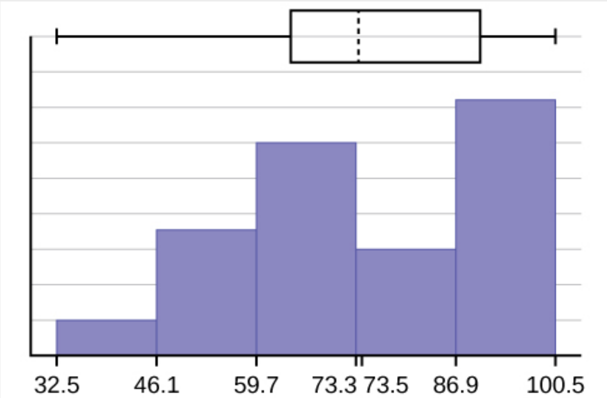
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution:

a. See [link]

b.
    i. The sample mean = 73.5
    ii. The sample standard deviation = 17.9
    iii. The median = 73
    iv. The first quartile = 61
    v. The third quartile = 90
    vi. *IQR* = 90 – 61 = 29

c. The *x*-axis goes from 32.5 to 100.5; *y*-axis goes from –2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is (100.5 – 32.5) divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, 32.5 + 13.6 = 46.1, 46.1 + 13.6 = 59.7, 59.7 + 13.6 = 73.3, 73.3 + 13.6 = 86.9, 86.9 + 13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 – 33 = 40) than the spread in the upper 50% (100 – 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Note:
Try It
Exercise:

Problem:

The following data show the different types of pet food stores in the area carry.
6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;
Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Solution:

$\mu = 9.3$

$s = 2.2$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$$

where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Example:
Find the standard deviation for the data in [link].

Class	Frequency, f	Midpoint, m	m^2	x^2	fm^2	Standard Deviation

Class	Frequency, *f*	Midpoint, *m*	m^2	$\bar{x}^2$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $\bar{x}$ = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-\bar{x})^2}{n-1}}$ where s_x = sample standard deviation, $\bar{x}$ = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

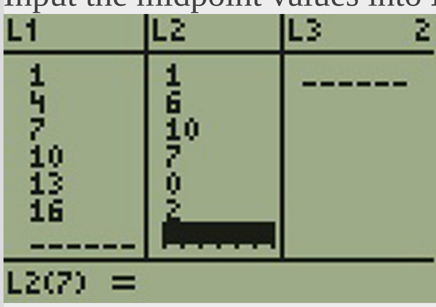
Note:
Try It
Find the standard deviation for the data from the previous example

Class	Frequency, *f*
0–2	1

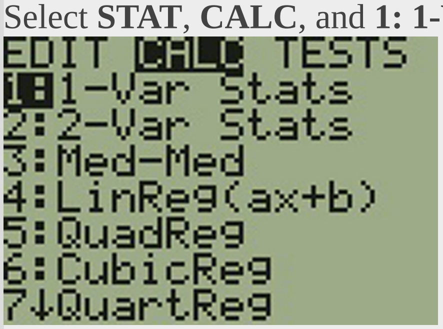
Class	Frequency, *f*
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

First, press the **STAT** key and select **1:Edit**

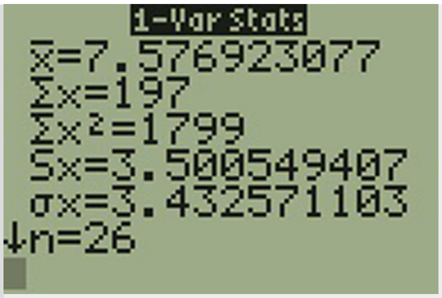
Input the midpoint values into **L1** and the frequencies into **L2**



Select **STAT**, **CALC**, and **1: 1-Var Stats**



Select **2nd** then **1** then , **2nd** then **2 Enter**



You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#ofSTDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = x + zs$	$z = \frac{x - x}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$z = \#\text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma}$

For John, $z = \#ofSTDEVs = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77-80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Note:

Try It
Exercise:

Problem:

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Solution:

For Angie: $z = \frac{26.2 - 27.2}{0.8} = -1.25$

For Beth: $z = \frac{27.3 - 30.1}{1.4} = -2$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.

- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

Data from Microsoft Bookshelf.

King, Bill.“Graphically Speaking.” Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Chapter Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-x)^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

Formula Review

$s_x = \sqrt{\frac{\sum fm^2}{n} - x^2}$ where $\begin{array}{l} s_x = \text{ sample standard deviation} \\ x = \text{sample mean} \end{array}$

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150
Exercise:

Problem:

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Solution:

$s = 34.5$

Exercise:

Problem: Find the value that is one standard deviation below the mean.

Exercise:

Problem:

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	**Batting Average**	**Team Batting Average**	**Team Standard Deviation**
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Solution:

For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$

For Karl: $z = \frac{0.177 - 0.189}{0.015} = -0.8$

Fredo's z-score of –0.67 is higher than Karl's z-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise:

Problem: Use [link] to find the value that is three standard deviations:

- **a**above the mean
- **b**below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise:

Problem:

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53

Daily Low Temperature	Frequency
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Solution:

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Homework

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1000 FTES
- median = 1,014 FTES
- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

Exercise:

Problem:

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution:

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

Exercise:

Problem: 75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

Exercise:

Problem: The population standard deviation = _____

Solution:

474 FTES

Exercise:

Problem:

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Exercise:

Problem: What is the *IQR*? What does the *IQR* represent?

Solution:

919

Exercise:

Problem: How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Exercise:

Problem:

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

Solution:

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

Exercise:

Problem:

What additional information is needed to construct a box plot for the FTES for 2005-2006 through 2010-2011 and a box plot for the FTES for 1976-1977 through 2004-2005?

Exercise:

Problem:

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

Solution:

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Exercise:

Problem:

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Exercise:

Problem:

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Exercise:

Problem:

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

Exercise:

Problem:

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in Table 14.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How "unusual" is the United States' obesity rate compared to the average rate? Explain.

Solution:

- $x = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that 23.32 + 12.95 = 36.27 is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not hav e an unusually high percentage of obese people.

Exercise:

Problem:

[link] gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

Bringing It Together

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4

# of movies	Frequency
4	1

a. Find the sample mean x.
b. Find the approximate sample standard deviation, s.

Solution:

a. 1.48
b. 1.12

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean x

b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40^{th} percentile.
k. Find the 90^{th} percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

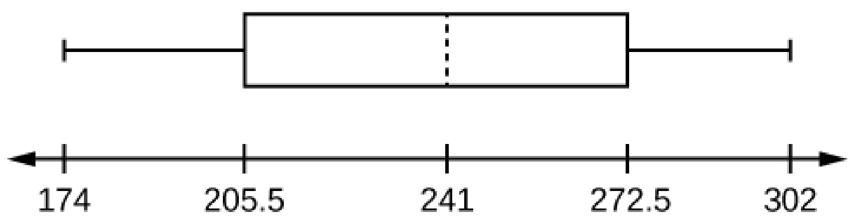
177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _______ to _______.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:

    i. the population mean, μ.
    ii. the population standard deviation, σ.
    iii. the weight that is two standard deviations below the mean.
    iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solution:

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5
e. 

f. 205.5, 272.5
g. sample
h. population
i.
    i. 236.34
    ii. 37.50
    iii. 161.34
    iv. 0.84 std. dev. below the mean
j. Young

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

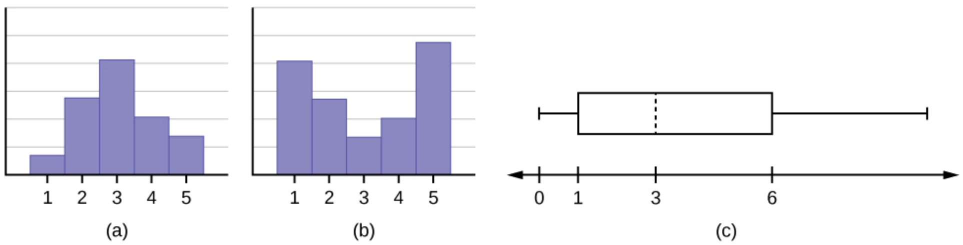
3 8–12 05–31–16 5–2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Refer to [link] determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

Solution:

a. True
b. True
c. True
d. False

Exercise:

Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.

b. Find the median, the first quartile, and the third quartile.
c. Find the 65^{th} percentile.
d. Find the 10^{th} percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _______ days to _______ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Solution:

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student’s solution.
c. mode
d. 8628.74
e. 6943.88
f. –0.09

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

x	**Frequency**
0	3
1	12
2	33
3	28
4	11

x	**Frequency**
5	9
6	4

Exercise:

Problem: The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Solution:

a

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the [link].

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - x$ where x is a value of the data and x is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Introduction
class="introduction"

Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.

- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Note:
Collaborative Exercise
Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. P(change) means the probability that a randomly chosen person in your class has change in his/her pocket or purse. P(bus) means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find P(change).
- Find P(bus).
- Find P(change AND bus). Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find P(change|bus). Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A

and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is {*HH*, *TH*, *HT*, *TT*} where *T* = tails and *H* = heads. The sample space has four outcomes. *A* = getting one head. There are two outcomes that meet this condition {*HT*, *TH*}, so $P(A) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event *E* = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces.

Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

"OR" Event:
An outcome is in the event *A* OR *B* if the outcome is in *A* or is in *B* or is in both *A* and *B*. For example, let *A* = {1, 2, 3, 4, 5} and *B* = {4, 5, 6, 7, 8}. *A* OR *B* = {1, 2, 3, 4, 5, 6, 7, 8}. Notice that 4 and 5 are NOT listed twice.

"AND" Event:
An outcome is in the event *A* AND *B* if the outcome is in both *A* and *B* at the same time. For example, let *A* and *B* be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then *A* AND *B* = {4, 5}.

The **complement** of event *A* is denoted *A′* (read "*A* prime"). *A′* consists of all outcomes that are **NOT** in *A*. Notice that $P(A) + P(A') = 1$. For example, let *S* = {1, 2, 3, 4, 5, 6} and let *A* = {1, 2, 3, 4}. Then, *A′* = {5, 6}. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$

The **conditional probability** of *A* given *B* is written $P(A|B)$. $P(A|B)$ is the probability that event *A* will occur given that the event *B* has already occurred. **A conditional reduces the sample space**. We calculate the probability of *A* from the reduced sample space *B*. The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \text{AND} B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space *S* = {1, 2, 3, 4, 5, 6}. Let *A* = face is 2 or 3 and *B* = face is even (2, 4, 6). To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space *B* = {2, 4, 6}. Then we divide that by the number of outcomes *B* (rather than *S*).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{\text{(the number of outcomes that are 2 or 3 and even in } S\text{)}}{6}}{\frac{\text{(the number of outcomes that are even in } S\text{)}}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example:
Exercise:

Problem:

The sample space S is the whole numbers starting at one and less than 20.

a. S = ________________________________

Let event A = the even numbers and event B = numbers greater than 13.

b. A = _______________________, B = _______________________
c. $P(A)$ = ______________, $P(B)$ = __________________
d. A AND B = ______________________, A OR B = __________________
e. $P(A$ AND $B)$ = __________, $P(A$ OR $B)$ = ______________
f. A' = ______________, $P(A')$ = ______________
g. $P(A) + P(A')$ = _____________
h. $P(A|B)$ = ____________, $P(B|A)$ = ______________; are the probabilities equal?

Solution:

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$
c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$
d. A AND $B = \{14,16,18\}$, A OR $B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$
f. $A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(A') = \frac{10}{19}$
g. $P(A) + P(A') = 1$ $\left(\frac{9}{19} + \frac{10}{19} = 1\right)$
h. $P(A|B) = \frac{P(A \text{AND} B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{AND} B)}{P(A)} = \frac{3}{9}$, No

Note:
Try It
Exercise:

Problem:

The sample space S is all the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. S = _________________________________

Let event A = the sum is even and event B = the first number is prime.

b. A = ________________________, B = ________________________
c. $P(A)$ = _______________, $P(B)$ = __________________
d. A AND B = _______________________, A OR B = __________________
e. $P(A \text{ AND } B)$ = __________, $P(A \text{ OR } B)$ = _______________
f. B' = _______________, $P(B')$ = _______________

g. $P(A) + P(A') =$ _____________

h. $P(A|B) =$ ____________, $P(B|A) =$ ______________; are the probabilities equal?

Solution:

a. S = {(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}

b. A = {(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)}

   B = {(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}

c. $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$

d. A AND B = {(2,2), (2,4), (3,1), (3,3)}

   A OR B = {(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}

e. $P(A \text{ AND } B) = \frac{1}{3}$, $P(A \text{ OR } B) = \frac{5}{6}$

f. B' = {(1,1), (1,2), (1,3), (1,4)}, $P(B') = \frac{1}{3}$

g. $P(B) + P(B') = 1$

h. $P(A|B) = \frac{P(A \text{AND} B)}{P(B)} = \frac{1}{2}$, $P(B|A) = \frac{P(A \text{AND} B)}{P(B)} = \frac{2}{3}$, No.

Example:
Exercise:

Problem:

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

a. Event T = the outcome is two.
b. Event A = the outcome is an even number.
c. Event B = the outcome is less than four.

d. The complement of *A*.
e. *A* GIVEN *B*
f. *B* GIVEN *A*
g. *A* AND *B*
h. *A* OR *B*
i. *A* OR *B′*
j. Event *N* = the outcome is a prime number.
k. Event *I* = the outcome is seven.

Solution:

a. $T = \{2\}$, $P(T) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
c. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$
d. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
e. $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
f. $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$
g. A AND $B = \{2\}$, $P(A \text{ AND } B) = \frac{1}{6}$
h. A OR $B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$
i. A OR $B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{2}{3}$
j. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

Example:
[link] describes the distribution of a random sample *S* of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Exercise:

Problem:

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

a. $P(M)$
b. $P(F)$
c. $P(R)$
d. $P(L)$
e. $P(M \text{ AND } R)$
f. $P(F \text{ AND } L)$
g. $P(M \text{ OR } F)$
h. $P(M \text{ OR } R)$
i. $P(F \text{ OR } L)$
j. $P(M')$
k. $P(R|M)$
l. $P(F|L)$
m. $P(L|F)$

Solution:

a. $P(M) = 0.52$
b. $P(F) = 0.48$
c. $P(R) = 0.87$
d. $P(L) = 0.13$
e. $P(M \text{ AND } R) = 0.43$
f. $P(F \text{ AND } L) = 0.04$

g. $P(M \text{ OR } F) = 1$
h. $P(M \text{ OR } R) = 0.96$
i. $P(F \text{ OR } L) = 0.57$
j. $P(M') = 0.48$
k. $P(R|M) = 0.8269$ (rounded to four decimal places)
l. $P(F|L) = 0.3077$ (rounded to four decimal places)
m. $P(L|F) = 0.0833$

References

"Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Chapter Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and *B* are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$P(A|B) = \frac{P(A \text{AND} B)}{P(B)}$

Exercise:

Problem:

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let *F* be the event that a student is female.
- Let *M* be the event that a student is male.
- Let *S* be the event that a student has short hair.
- Let *L* be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Solution:

a. $P(L') = P(S)$
b. $P(M \text{ OR } S)$
c. $P(F \text{ AND } L)$
d. $P(M|L)$

e. $P(L|M)$
f. $P(S|F)$
g. $P(F|L)$
h. $P(F \text{ OR } L)$
i. $P(M \text{ AND } S)$
j. $P(F)$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.
Let H = the event of getting a hat.
Let N = the event of getting a noisemaker.
Let F = the event of getting a finger trap.
Let C = the event of getting a bag of confetti.

Exercise:

Problem: Find $P(H)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(C)$.

Solution:

$P(C) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.
Let B = the event of getting a blue jelly bean
Let G = the event of getting a green jelly bean.
Let O = the event of getting an orange jelly bean.
Let P = the event of getting a purple jelly bean.
Let R = the event of getting a red jelly bean.
Let Y = the event of getting a yellow jelly bean.

Exercise:

Problem: Find $P(B)$.

Exercise:

Problem: Find $P(G)$.

Solution:

$P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise:

Problem: Find $P(P)$.

Exercise:

Problem: Find $P(R)$.

Solution:

$P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise:

Problem: Find $P(Y)$.

Exercise:

Problem: Find $P(O)$.

Solution:

$P(O) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).
Let A = the event that a country is in Asia.
Let E = the event that a country is in Europe.
Let F = the event that a country is in Africa.
Let N = the event that a country is in North America.
Let O = the event that a country is in Oceania.
Let S = the event that a country is in South America.

Exercise:

Problem: Find $P(A)$.

Exercise:

Problem: Find $P(E)$.

Solution:

$P(E) = \frac{47}{194} = 0.24$

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$P(N) = \frac{23}{194} = 0.12$

Exercise:

Problem: Find $P(O)$.

Exercise:

Problem: Find $P(S)$.

Solution:

$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise:

Problem:

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise:

Problem:

What is the probability of drawing a club in a standard deck of 52 cards?

Solution:

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise:

Problem:

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise:

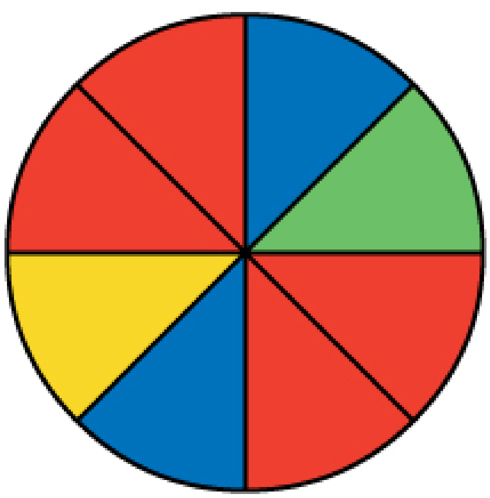
Problem:

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Solution:

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.



Let B = the event of landing on blue.
Let R = the event of landing on red.

Let G = the event of landing on green.
Let Y = the event of landing on yellow.
Exercise:

Problem: If you land on Y, you get the biggest prize. Find $P(Y)$.

Exercise:

Problem: If you land on red, you don't get a prize. What is $P(R)$?

Solution:

$$P(R) = \frac{4}{8} = 0.5$$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.
Let I = the event that a player in an infielder.
Let O = the event that a player is an outfielder.
Let H = the event that a player is a great hitter.
Let N = the event that a player is not a great hitter.
Exercise:

Problem:

Write the symbols for the probability that a player is not an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Solution:

$P(O \text{ OR } H)$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Solution:

$P(H|I)$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Solution:

$P(N|O)$

Exercise:

Problem:

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Solution:

$P(I \text{ OR } N)$

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder.

Solution:

$P(I)$

Exercise:

Problem: What is the word for the set of all possible outcomes?

Exercise:

Problem: What is conditional probability?

Solution:

The likelihood that an event will occur given that another event has already occurred.

Exercise:

Problem:

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book
Let F = event that book is fiction
Let N = event that book is nonfiction
What is the sample space?

Exercise:

Problem:

What is the sum of the probabilities of an event and its complement?

Solution:

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

Exercise:

Problem: What does $P(E|M)$ mean in words?

Exercise:

Problem: What does $P(E \text{ OR } M)$ mean in words?

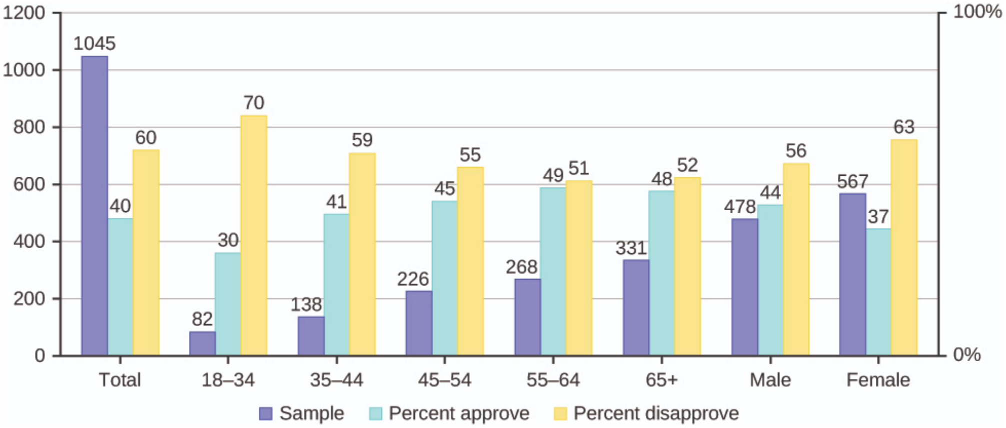
Solution:

the probability of landing on an even number or a multiple of three

Homework

Exercise:

Problem:



The graph in [link] displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find *P*(Approve|Male).
i. Out of the age groups, what percent are more than 44 years old?
j. Find *P*(Approve|Age < 35).

Exercise:

Problem:

Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

Solution:

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

Glossary

Conditional Probability
: the likelihood that an event will occur given that another event has already occurred

Equally Likely
: Each outcome of an experiment has the same probability.

Event
: a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by *S*. An event is an arbitrary subset in *S*. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as *A*, *B*, *C*, and so on.

Experiment
: a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then:

- $0 \leq P(A) \leq 1$
- If A and B are any two mutually exclusive events, then $P(A \text{ OR } B) = P(A) + P(\text{B})$.
- $P(S) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of A GIVEN B

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The Or Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

Independent and Mutually Exclusive Events

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Two events *A* and *B* are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

Example:
You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit.
a. Sampling with replacement:
Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.
b. Sampling without replacement:
Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

Note:
Try It
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with

or without replacement?
b. Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?

Solution:

a. With replacement
b. No

Example:
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are QS, $1D$, $1C$, QD.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are KH, $7D$, $6D$, KH.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Solution:

a. Without replacement; b. With replacement

Note:
Try It
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. *QS*, 1*D*, 1*C*, *QD*
b. *KH*, 7*D*, 6*D*, *KH*
c. *QS*, 7*D*, 6*D*, *KS*

Solution:

without replacement: 1. Possible; 2. Impossible, 3. Possible

with replacement: 1. Possible; 2. Possible, 3. Possible

Mutually Exclusive Events

A and *B* are **mutually exclusive** events if they cannot occur at the same time. This means that *A* and *B* do not share any outcomes and *P*(*A* AND *B*) = 0.

For example, suppose the sample space *S* = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let *A* = {1, 2, 3, 4, 5}, *B* = {4, 5, 6, 7, 8}, and *C* = {7, 9}. *A* AND *B* = {4, 5}. *P*(*A* AND *B*) = $\frac{2}{10}$ and is not equal to zero. Therefore, *A* and *B* are not mutually exclusive. *A* and *C* do not have any numbers in common so *P*(*A* AND *C*) = 0. Therefore, *A* and *C* are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example:
Flip two fair coins. (This is an experiment.)
The sample space is $\{HH, HT, TH, TT\}$ where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as $\{HH, HT, TH\}$. The outcome HH shows zero tails. HT and TH each show one tail.
- Let B = the event of getting all tails. B can be written as $\{TT\}$. B is the **complement** of A, so $B = A'$. Also, $P(A) + P(B) = P(A) + P(A') = 1$.
- The probabilities for A and for B are $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{4}$.
- Let C = the event of getting all heads. $C = \{HH\}$. Since $B = \{TT\}$, $P(B \text{ AND } C) = 0$. B and C are mutually exclusive. (B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. $D = \{TT\}$. $P(D) = \frac{1}{4}$
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $E = \{HT, HH\}$. $P(E) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let F = event of getting at least one tail in two flips. $F = \{HT, TH, TT\}$. $P(F) = \frac{3}{4}$

Note:
Try It
Exercise:

Problem:

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Solution:
Try It Solutions

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is {*BB*, *BR*, *RB*, *RR*}.

Event A = Getting at least one black card = {*BB*, *BR*, *RB*}

$P(A) = \frac{3}{4} = 0.75$

Example:
Exercise:

Problem: Flip two fair coins. Find the probabilities of the events.

a. Let F = the event of getting at most one tail (zero or one tail).
b. Let G = the event of getting two faces that are the same.
c. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are F and G mutually exclusive?
e. Let J = the event of getting all tails. Are J and H mutually exclusive?

Solution:

Look at the sample space in [link].

a. Zero (0) or one (1) tails occur when the outcomes *HH*, *TH*, *HT* show up. $P(F) = \frac{3}{4}$

b. Two faces are the same if *HH* or *TT* show up. $P(G) = \frac{2}{4}$
c. A head on the first flip followed by a head or tail on the second flip occurs when *HH* or *HT* show up. $P(H) = \frac{2}{4}$
d. *F* and *G* share *HH* so $P(F \text{ AND } G)$ is not equal to zero (0). *F* and *G* are not mutually exclusive.
e. Getting all tails occurs when tails shows up on both coins (*TT*). *H*'s outcomes are *HH* and *HT*.

J and *H* have nothing in common so $P(J \text{ AND } H) = 0$. *J* and *H* are mutually exclusive.

Note:
Try It
Exercise:

Problem:

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let *F* = the event of getting the white ball twice.
b. Let *G* = the event of getting two balls of different colors.
c. Let *H* = the event of getting white on the first pick.
d. Are *F* and *G* mutually exclusive?
e. Are *G* and *H* mutually exclusive?

Solution:

a. $P(F) = \frac{1}{4}$
b. $P(G) = \frac{1}{2}$
c. $P(H) = \frac{1}{2}$
d. Yes
e. No

Example:

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Exercise:

Problem:

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Solution:

No. C = {3, 5} and E = {1, 2, 3, 4}. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$.

Note:
Try It
Exercise:

Problem:

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A)$ = 0.4 and $P(B)$ = 0.2. $P(A \text{ AND } B)$ = 0.08. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Solution:

$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{0.08}{0.2} = 0.4 = P(A)$

The events are independent because $P(A|B) = P(A)$.

Example:
Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(G)$ = 0.6, $P(H)$ = 0.5, and $P(G \text{ AND } H)$ = 0.3. Are G and H independent?
If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

Note:
NOTE

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

Exercise:

Problem: a. Show that $P(G|H) = P(G)$.

Solution:

$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$

Exercise:

Problem: b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution:

$P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Note:
Try It
Exercise:

Problem:

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

Solution:

Event G and $O = \{G1, G3\}$

$P(G \text{ and } O) = \frac{2}{10} = 0.2$

Example:
Exercise:

Problem:

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(D|C)$?

Solution:

a. Yes, because $P(C|D) = P(C)$.
b. No, because $P(C \text{ AND } D)$ is not equal to zero.
c. $P(D|C) = \frac{P(C \text{ AND } D)}{P(C)} = \frac{0.225}{0.75} = 0.3$

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

a. Find $P(B|D)$.
b. Find $P(D|B)$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Solution:

a. $P(B|D) = 0.6667$
b. $P(D|B) = 0.5$
c. No
d. No

Example:

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.
Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.
The sample space S = $R1$, $R2$, $R3$, $B1$, $B2$, $B3$, $B4$, $B5$. S has eight outcomes.

- $P(R) = \frac{3}{8}$. $P(B) = \frac{5}{8}$. $P(R \text{ AND } B) = 0$. (You cannot draw one card that is both red and blue.)
- $P(E) = \frac{3}{8}$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(E|B) = \frac{2}{5}$. (There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(B|E) = \frac{2}{3}$. (There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.
- Let G = card with a number greater than 3. $G = \{B4, B5\}$. $P(G) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $H = \{B1, B2, B3, B4\}$. $P(G|H) = \frac{1}{4}$. (The only card in H that has a number greater than three is $B4$.) Since $\frac{2}{8} = \frac{1}{4}$, $P(G) = P(G|H)$, which means that G and H are independent.

Note:
Try It
Exercise:

Problem: In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.

- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.
Let B be the event that a fan is wearing blue.
Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Solution:

$P(B|A) = 0.67$

$P(B) = 0.25$

So $P(B)$ does not equal $P(B|A)$ which means that B and A are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND } A) = 0.20$, not 0.

Example:
In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

Note:
NOTE

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(F|L)$ yet, so you cannot use the second condition.

Solution 1
Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.

Solution 2
Check whether $P(L|F)$ equals $P(L)$. We are given that $P(L|F) = 0.75$, but $P(L) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results
The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Note:
Try It
Exercise:

Problem:

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.56$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

Solution:

Because $P(I \text{ AND } F) = 0$,

$P(I \text{ OR } F) = P(I) + P(F) - P(I \text{ AND } F) = 0.44 + 0.56 - 0 = 1$

Example:
Exercise:

Problem:

a. Toss one fair coin (the coin has two sides, *H* and *T*). The outcomes are _________. Count the outcomes. There are _____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are __________________. Count the outcomes. There are ___ outcomes.
c. Multiply the two numbers of outcomes. The answer is ________.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in part c. is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are *H*1 and *T*6.)
e. Event *A* = heads (*H*) on the coin followed by an even number (2, 4, 6) on the die.
A = {___________________}. Find $P(A)$.
f. Event *B* = heads on the coin followed by a three on the die. *B* = {_________}. Find $P(B)$.
g. Are *A* and *B* mutually exclusive? (Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then *A* and *B* are mutually exclusive.)
h. Are *A* and *B* independent? (Hint: Is $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then *A* and *B* are independent. If not, then they are dependent).

Solution:

a. *H* and *T*; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. *T*1, *T*2, *T*3, *T*4, *T*5, *T*6, *H*1, *H*2, *H*3, *H*4, *H*5, *H*6

e. $A = \{H2, H4, H6\}$; $P(A) = \frac{3}{12}$
f. $B = \{H3\}$; $P(B) = \frac{1}{12}$
g. Yes, because $P(A \text{ AND } B) = 0$
h. $P(A \text{ AND } B) = 0.P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ AND } B)$ does not equal $P(A)P(B)$, so A and B are dependent.

Note:
Try It
Exercise:

Problem:

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

a. Compute $P(T)$.
b. Compute $P(T|F)$.
c. Are T and F independent?.
d. Are F and S mutually exclusive?
e. Are F and S independent?

Solution:

a. $P(T) = \frac{1}{4}$
b. $P(T|F) = \frac{1}{2}$
c. No
d. No
e. Yes

References

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Chapter Review

Two events *A* and *B* are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

If A and B are independent, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

Exercise:

Problem:

E and F are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E|F)$.

Exercise:

Problem: J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

Solution:

$P(J) = 0.3$

Exercise:

Problem:

U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

a. $P(U \text{ AND } V) =$
b. $P(U|V) =$
c. $P(U \text{ OR } V) =$

Exercise:

Problem:

Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

Solution:

$P(Q \text{ AND } R) = P(Q)P(R)$

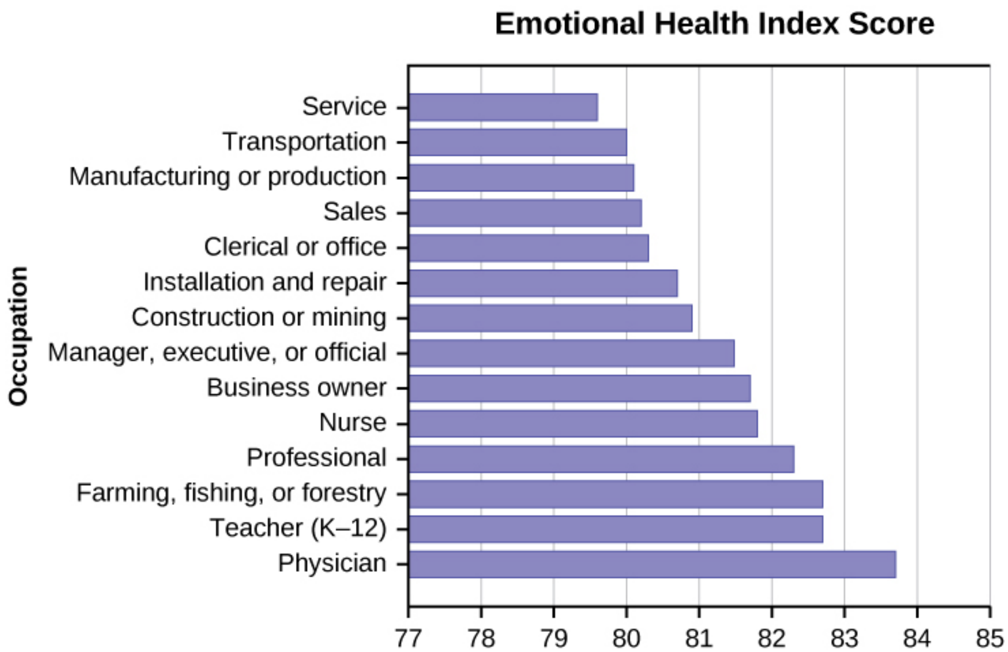
$0.1 = (0.4)P(R)$

$P(R) = 0.25$

Homework

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of

employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.



Exercise:

Problem:

Find the probability that an Emotional Health Index Score is 82.7.

Exercise:

Problem:

Find the probability that an Emotional Health Index Score is 81.0.

Solution:

0

Exercise:

Problem:

Find the probability that an Emotional Health Index Score is more than 81?

Exercise:

Problem:

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

Solution:

0.3571

Exercise:

Problem:

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

Exercise:

Problem:

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

Solution:

0.2142

Exercise:

Problem:

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

Exercise:

Problem: What occupation has the highest emotional index score?

Solution:

Physician (83.7)

Exercise:

Problem: What occupation has the lowest emotional index score?

Exercise:

Problem: What is the range of the data?

Solution:

83.7 − 79.6 = 4.1

Exercise:

Problem: Compute the average EHIS.

Exercise:

Problem:

If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

Solution:

$P(\text{Occupation} < 81.3) = 0.5$

Bringing It Together

Exercise:

Problem:

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into [link].

Shirt#	**≤ 210**	**211–250**	**251–290**	**290≤**
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about *P*(Shirt# 1–33|≤ 210 pounds)?

Exercise:

Problem:

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let *C* = a man develops cancer in his lifetime and *P* = man has at least one false positive.

a. $P(C)$ = _______
b. $P(P|C)$ = _______
c. $P(P|C')$ = _______
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Solution:

a. $P(C) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise:

Problem:

Given events G and H: $P(G) = 0.43$; $P(H) = 0.26$; $P(H \text{ AND } G) = 0.14$

a. Find $P(H \text{ OR } G)$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

Exercise:

Problem:

Given events J and K: $P(J) = 0.18$; $P(K) = 0.37$; $P(J \text{ OR } K) = 0.45$

a. Find $P(J \text{ AND } K)$.
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J OR K).

Solution:

a. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; $0.45 = 0.18 + 0.37 - P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K) = 0.10$

b. $P(\text{NOT } (J \text{ AND } K)) = 1 - P(J \text{ AND } K) = 1 - 0.10 = 0.90$

c. $P(\text{NOT } (J \text{ OR } K)) = 1 - P(J \text{ OR } K) = 1 - 0.45 = 0.55$

Glossary

Dependent Events
: If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement
: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement
: When sampling is done without replacement, each member of a population may be chosen only once.

The Conditional Probability of One Event Given Another Event
: $P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The OR of Two Events
: An outcome is in the event A OR B if the outcome is in A, is in B, or is in both A and B.

Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a **sample space**, then: $P(A \text{ AND } B) = P(B)P(A|B)$.

This rule may also be written as: $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are **independent**, then $P(A|B) = P(A)$. Then $P(A \text{ AND } B) = P(A|B)P(B)$ becomes $P(A \text{ AND } B) = P(A)P(B)$.

The Addition Rule

If A and B are defined on a sample space, then: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$.

If A and B are **mutually exclusive**, then $P(A \text{ AND } B) = 0$. Then $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$ becomes $P(A \text{ OR } B) = P(A) + P(B)$.

Example:
Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses A is $P(A) = 0.6$ and the probability that he chooses B is $P(B) = 0.35$.
- $P(A \text{ AND } B) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \text{ OR } B) = P(A) + P(B) = 0.6 + 0.35 = 0.95$. Note that the

probability that he does not choose to go anywhere on vacation must be 0.05.

Example:
Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(B) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

Exercise:

Problem: a. What is the probability that he makes both goals?

Solution:

a. The problem is asking you to find $P(A \text{ AND } B) = P(B \text{ AND } A)$. Since $P(B|A) = 0.90$: $P(B \text{ AND } A) = P(B|A)\ P(A) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

Exercise:

Problem:

b. What is the probability that Carlos makes either the first goal or the second goal?

Solution:

b. The problem is asking you to find $P(A \text{ OR } B)$.

$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) = 0.65 + 0.65 - 0.585 = 0.715$

Carlos makes either the first goal or the second goal with probability 0.715.

Exercise:

Problem: c. Are A and B independent?

Solution:

c. No, they are not, because $P(B \text{ AND } A) = 0.585$.

$P(B)P(A) = (0.65)(0.65) = 0.423$

$0.423 \neq 0.585 = P(B \text{ AND } A)$

So, $P(B \text{ AND } A)$ is **not** equal to $P(B)P(A)$.

Exercise:

Problem: d. Are A and B mutually exclusive?

Solution:

d. No, they are not because $P(A \text{ and } B) = 0.585$.

To be mutually exclusive, $P(A \text{ AND } B)$ must equal zero.

Note:
Try It
Exercise:

Problem:

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(C) = 0.75$. D = the event Helen makes the second shot. $P(D) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Solution:

$P(D|C) = 0.85$

$P(C \text{ AND } D) = P(D \text{ AND } C)$
$P(D \text{ AND } C) = P(D|C)P(C) = (0.85)(0.75) = 0.6375$
Helen makes the first and second free throws with probability 0.6375.

Example:
A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

Exercise:

Problem:

a. What is the probability that the member is a novice swimmer?

Solution:

a. $\frac{28}{150}$

Exercise:

Problem:

b. What is the probability that the member practices four times a week?

Solution:

b. $\frac{80}{150}$

Exercise:

Problem:

c. What is the probability that the member is an advanced swimmer and practices four times a week?

Solution:

c. $\frac{40}{150}$

Exercise:

Problem:

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

Solution:

d. P(advanced AND intermediate) = 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Exercise:

Problem:

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution:

e. No, these are not independent events.
P(novice AND practices four times per week) = 0.0667
P(novice)P(practices four times per week) = 0.0996
0.0667 ≠ 0.0996

Note:
Try It
Exercise:

Problem:

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Solution:

$$P = \frac{200-140-40}{200} = \frac{20}{200} = 0.1$$

Example:
Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, $M|S$ = math given speech

Exercise:

Problem:

a. What is the probability that Felicity enrolls in math and speech? Find $P(M \text{ AND } S) = P(M|S)P(S)$.
b. What is the probability that Felicity enrolls in math or speech classes?
Find $P(M \text{ OR } S) = P(M) + P(S) - P(M \text{ AND } S)$.
c. Are M and S independent? Is $P(M|S) = P(M)$?
d. Are M and S mutually exclusive? Is $P(M \text{ AND } S) = 0$?

Solution:

a. 0.1625, b. 0.6875, c. No, d. No

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

a. Find $P(B \text{ AND } D)$.
b. Find $P(B \text{ OR } D)$.

Solution:

a. $P(B \text{ AND } D) = P(D|B)P(B) = (0.5)(0.4) = 0.20$.
b. $P(B \text{ OR } D) = P(B) + P(D) - P(B \text{ AND } D) = 0.40 + 0.30 - 0.20 = 0.50$

Example:
Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

Exercise:

Problem:

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

Solution:

a. $P(B) = 0.143$; $P(N) = 0.85$

Exercise:

Problem:

b. Given that the woman has breast cancer, what is the probability that she tests negative?

Solution:

b. $P(N|B) = 0.02$

Exercise:

Problem:

c. What is the probability that the woman has breast cancer AND tests negative?

Solution:

c. $P(B \text{ AND } N) = P(B)P(N|B) = (0.143)(0.02) = 0.0029$

Exercise:

Problem:

d. What is the probability that the woman has breast cancer or tests negative?

Solution:

d. $P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = 0.143 + 0.85 - 0.0029 = 0.9901$

Exercise:

Problem:

e. Are having breast cancer and testing negative independent events?

Solution:

e. No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

Exercise:

Problem:

f. Are having breast cancer and testing negative mutually exclusive?

Solution:

f. No. $P(B \text{ AND } N) = 0.0029$. For B and N to be mutually exclusive, $P(B \text{ AND } N)$ must be zero.

Note:
Try It
Exercise:

Problem:

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Solution:

Let A = student is a senior going to college.

Let B = student plays sports.

$P(B) = \frac{140}{200}$

$P(B|A) = \frac{50}{140}$

$P(A \text{ AND } B) = P(B|A)P(A)$

$P(A \text{ AND } B) = \left(\frac{140}{200}\right)\left(\frac{50}{140}\right) = \frac{1}{4}$

Example:
Exercise:

Problem: Refer to the information in [link]. P = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P|B) = 1 - P(N|B)$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(B \text{ AND } P) = P(P|B)P(B)$.

c. What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$.

Solution:

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

a. Find $P(B')$.
b. Find $P(D \text{ AND } B)$.
c. Find $P(B|D)$.
d. Find $P(D \text{ AND } B')$.
e. Find $P(D|B')$.

Solution:

a. $P(B') = 0.60$
b. $P(D \text{ AND } B) = P(D|B)P(B) = 0.20$
c. $P(B|D) = \frac{P(B \text{ AND } D)}{P(D)} = \frac{(0.20)}{(0.30)} = 0.66$
d. $P(D \text{ AND } B') = P(D) - P(D \text{ AND } B) = 0.30 - 0.20 = 0.10$
e. $P(D|B') = P(D \text{ AND } B')P(B') = (P(D) - P(D \text{ AND } B))(0.60) = (0.10)(0.60) = 0.06$

References

DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf (accessed May 2, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Chapter Review

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(A \text{ AND } B) = P(A|B)P(B)$

The addition rule: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.
Exercise:

Problem: Find $P(C)$.

Exercise:

Problem: Find $P(L)$.

Solution:

0.376

Exercise:

Problem: Find $P(C|L)$.

Exercise:

Problem: In words, what is $C|L$?

Solution:

$C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Exercise:

Problem: Find $P(L \text{ AND } C)$.

Exercise:

Problem: In words, what is L AND C?

Solution:

L AND C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Exercise:

Problem: Are L and C independent events? Show why or why not.

Exercise:

Problem: Find $P(L \text{ OR } C)$.

Solution:

0.6492

Exercise:

Problem: In words, what is L OR C?

Exercise:

Problem:

Are L and C mutually exclusive events? Show why or why not.

Solution:

No, because $P(L \text{ AND } C)$ does not equal 0.

Homework

Exercise:

Problem:

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- C = California registered voters who support same-sex marriage.

- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18 to 39 years old.

a. Find $P(C)$.
b. Find $P(B)$.
c. Find $P(C|A)$.
d. Find $P(B|C)$.
e. In words, what is $C|A$?
f. In words, what is $B|C$?
g. Find $P(C \text{ AND } B)$.
h. In words, what is C AND B?
i. Find $P(C \text{ OR } B)$.
j. Are C and B mutually exclusive events? Show why or why not.

Exercise:

Problem:

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

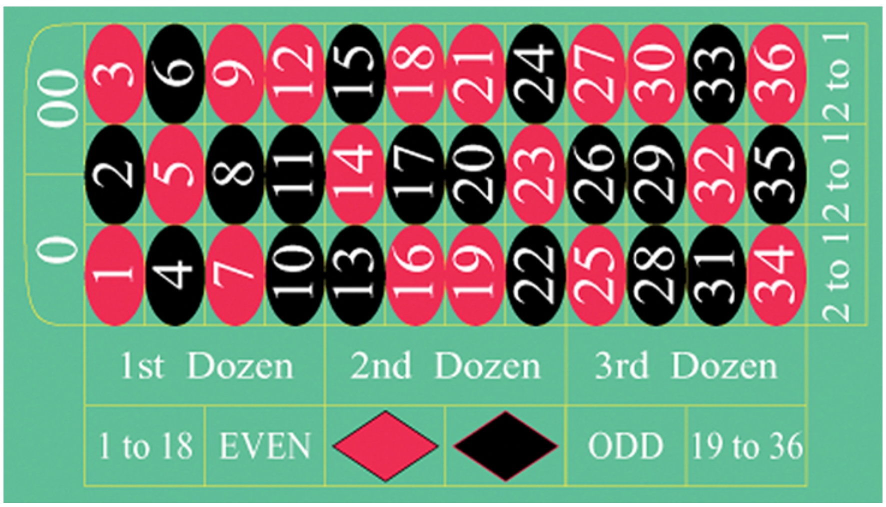
a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?

e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

Solution:

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.



(credit: film8ker/wikibooks)

Exercise:

Problem:

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find *P*(red).
c. You bet on -1st 12- (1st Dozen). Find *P*(-1st 12-).
d. You bet on an even number. Find *P*(even number).
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Exercise:

Problem:

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

Solution:

a. *P*(Betting on two line that touch each other on the table) = $\frac{6}{38}$
b. *P*(Betting on three numbers in a line) = $\frac{3}{38}$
c. *P*(Bettting on one number) = $\frac{1}{38}$
d. *P*(Betting on four number that touch each other to form a square) = $\frac{4}{38}$
e. *P*(Betting on two number that touch each other on the table) = $\frac{2}{38}$
f. *P*(Betting on 0-00-1-2-3) = $\frac{5}{38}$
g. *P*(Betting on 0-1-2; or 0-00-2; or 00-2-3) = $\frac{3}{38}$

Exercise:

Problem:

Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

Exercise:

Problem:

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. $P(G)$ = _____
c. $P(G|E)$ = _____
d. $P(G \text{ AND } E)$ = _____
e. $P(G \text{ OR } E)$ = _____
f. Are G and E mutually exclusive? Justify your answer numerically.

Solution:

a. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
b. $\frac{5}{8}$
c. $\frac{2}{3}$

d. $\frac{2}{8}$
e. $\frac{6}{8}$
f. No, because $P(G \text{ AND } E)$ does not equal 0.

Exercise:

Problem: Roll two fair dice separately. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
d. In words, explain what "$P(A|B)$" represents. Find $P(A|B)$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

Exercise:

Problem:

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Solution:

> **Note:**
> **NOTE**
> The coin toss is independent of the card picked first.

a. {(G,H) (G,T) (B,H) (B,T) (R,H) (R,T)}
b. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \text{ AND } B) = 0$
d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(A \text{ AND } C) = P(A) = \frac{3}{20}$

Exercise:

Problem:

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Exercise:

Problem:

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

Solution:

a. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
b. $\frac{4}{8}$
c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \text{ AND } B) = 0$.

Exercise:

Problem: Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Let $P(C|D) = 0.6$.

a. Find $P(C \text{ AND } D)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(C \text{ OR } D)$.
e. Find $P(D|C)$.

Exercise:

Problem: Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y \text{ AND } Z)$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(Z)$ if $P(Y \text{ OR } Z) = 0.71$ and $P(Y) = 0.42$.

Solution:

a. If Y and Z are independent, then $P(Y \text{ AND } Z) = P(Y)P(Z)$, so $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y)P(Z)$.
b. 0.5

Exercise:

Problem: G and H are mutually exclusive events. $P(G) = 0.5$ $P(H) = 0.3$

a. Explain why the following statement MUST be false: $P(H|G) = 0.4$.
b. Find $P(H \text{ OR } G)$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Exercise:

Problem:

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E' = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(E')$ =	i. 0.8043
b. $P(E)$ =	ii. 0.623
c. $P(S \text{ and } E')$ =	iii. 0.1957
d. $P(S\|E')$ =	iv. 0.1219

Solution:

iii i iv ii

Exercise:

Problem:

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Exercise:

Problem:

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.
c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

Solution:

a. $P(R) = 0.44$
b. $P(R|E) = 0.56$
c. $P(R|O) = 0.31$
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

Exercise:

Problem:

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because P(hit by Hank Aaron|hit is a double) = P(hit by Hank Aaron)
b. No, because P(hit by Hank Aaron|hit is a double) ≠ P(hit is a double)
c. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron)

d. Yes, because P(hit is by Hank Aaron|hit is a double) = P(hit is a double)

Exercise:

Problem:

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

Solution:

a. P(type O OR Rh-) = P(type O) + P(Rh-) - P(type O AND Rh-)

   0.52 = 0.43 + 0.15 - P(type O AND Rh-); solve to find P(type O AND Rh-) = 0.06

   6% of people have type O, Rh- blood

b. P(NOT(type O AND Rh-)) = 1 - P(type O AND Rh-) = 1 - 0.06 = 0.94

   94% of people do not have type O, Rh- blood

Exercise:

Problem:

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

a. Find the probability that a course has a final exam or a research project.
b. Find the probability that a course has NEITHER of these two requirements.

Exercise:

Problem:

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

Solution:

a. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
b. $P(C \text{ OR } N) = P(C) + P(N) - P(C \text{ AND } N) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \text{ OR } N) = 1 - 0.40 = 0.60$

Exercise:

Problem:

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student

a. Find $P(D \text{ AND } E)$.
b. Find $P(E|D)$.
c. Find $P(D \text{ OR } E)$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \text{ AND } B) = P(A)P(B)$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \text{ AND } B) = 0$.

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example:
Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.
Calculate the following probabilities using the table.

a. Find *P*(Driver is a cell phone user).
b. Find *P*(driver had no violation in the last year).
c. Find *P*(Driver had no violation in the last year AND was a cell phone user).
d. Find *P*(Driver is a cell phone user OR driver had no violation in the last year).
e. Find *P*(Driver is a cell phone user GIVEN driver had a violation in the last year).
f. Find *P*(Driver had no violation last year GIVEN driver was not a cell phone user)

Solutions:

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$

b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

c. $\frac{280}{755}$

d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$

e. $\frac{25}{70}$ (The sample space is reduced to the number of drivers who had a violation.)

f. $\frac{405}{450}$ (The sample space is reduced to the number of drivers who were not cell phone users.)

Note:
Try it
Exercise:

Problem:

[link] shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is *P*(athlete stretches before exercising)?
b. What is *P*(athlete stretches before exercising|no injury in the last year)?

Solution:

a. *P*(athlete stretches before exercising) = $\frac{350}{800}$ = 0.4375
b. *P*(athlete stretches before exercising|no injury in the last year) = $\frac{295}{514}$ = 0.5739

Example:
[link] shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Male	___	___	14	55
Total	___	41	___	___

Hiking Area Preference

Exercise:

Problem: a. Complete the table.

Solution:

a.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Hiking Area Preference

Exercise:

Problem: b. Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Solution:

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = 0.18$
2. $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$

$P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C are not independent.

Exercise:

Problem:

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: $P(\_\_\_|\_\_\_) = \_\_\_$.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Solution:

c.

1. The word 'given' tells you that this is a conditional.
2. $P(M|L) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

Exercise:

Problem:

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Solution:

d.

1. $P(F) = \frac{45}{100}$
2. $P(P) = \frac{25}{100}$
3. $P(F \text{ AND } P) = \frac{11}{100}$
4. $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Note:
Try It
Exercise:

Problem:

[link] shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Solution:

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$

$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$

$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example:
Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Caught or Not	Door One	Door Two	Door Three	Total

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

Door Choice

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Caught)
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Not Caught)

Verify the remaining entries.

Exercise:

Problem:

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Solution:

a.

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

Door Choice

Exercise:

Problem: b. What is the probability that Alissa does not catch Muddy?

Solution:

b. $\frac{41}{60}$

Exercise:

Problem:

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution:

c. $\frac{9}{19}$

Example:

[link] contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Year	**Robbery**	**Burglary**	**Rape**	**Vehicle**	**Total**
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Exercise:

Problem: TOTAL each column and each row. Total data = 4,520.7

a. Find P(2009 AND Robbery).
b. Find P(2010 AND Burglary).
c. Find P(2010 OR Burglary).
d. Find P(2011|Rape).
e. Find P(Vehicle|2008).

Solution:

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Note:
Try It
Exercise:

Problem:

[link] relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Solution:

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. P(Tall) = $\frac{50}{205}$ = 0.244
c. P(Obese AND Tall) = $\frac{18}{205}$ = 0.088
d. P(Tall|Obese) = $\frac{18}{60}$ = 0.3
e. P(Obese|Tall) = $\frac{18}{50}$ = 0.36
f. P(Tall AND Underweight = $\frac{12}{205}$ = 0.0585
g. No. P(Tall) does not equal P(Tall|Obese).

References

"Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/blood-types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

"Human Blood Types." Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Samuel, T. M. "Strange Facts about RH Negative Blood." eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

"United States: Uniform Crime Report – State Statistics from 1960–2011." The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Chapter Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilites that have multiple dependent variables.

Use the following information to answer the next four exercises. [link] shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Exercise:

Problem: Find *P*(musician is a female).

Exercise:

Problem: Find *P*(musician is a male AND had private instruction).

Solution:

P(musician is a male AND had private instruction) = $\frac{15}{130}$ = $\frac{3}{26}$ = 0.12

Exercise:

Problem: Find *P*(musician is a female OR is self taught).

Exercise:

Problem:

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

Solution:

P(being a female musician AND learning music in school) = $\frac{38}{130}$ = $\frac{19}{65}$ = 0.29

P(being a female musician)*P*(learning music in school) = $\left(\frac{72}{130}\right)\left(\frac{62}{130}\right)$ =

$$\frac{4,464}{16,900}$$

= $\frac{1,116}{4,225}$ = 0.26

No, they are not independent because *P*(being a female musician AND learning music in school) is not equal to *P*(being a female musician)*P*(learning music in school).

Bringing It Together

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378

Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

Exercise:

Problem:

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Level	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
1–10						
11–20						
21–30						
31+						
TOTALS						

Smoking Levels by Ethnicity

Exercise:

Problem:

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Solution:

$\frac{35{,}065}{100{,}450}$

Exercise:

Problem: Find the probability that the person was Latino.

Exercise:

Problem:

In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Also, find the probability.

Solution:

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

Exercise:

Problem:

In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Also, find the probability.

Exercise:

Problem:

In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Also, find the probability.

Solution:

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

Exercise:

Problem: Prove that smoking level/day and ethnicity are dependent events.

Homework

Use the information in the [link] to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Exercise:

Problem: What is the probability that a randomly selected senator has an "Other" affiliation?

Solution:

0

Exercise:

Problem:

What is the probability that a randomly selected senator is up for reelection in November 2016?

Exercise:

Problem:

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

Solution:

$\frac{10}{67}$

Exercise:

Problem:

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Exercise:

Problem:

Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

Solution:

$\frac{10}{34}$

Exercise:

Problem:

Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

Exercise:

Problem: The events “Republican” and “Up for reelection in 2016” are ________

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Solution:

d

Exercise:

Problem: The events “Other” and “Up for reelection in November 2016” are _________

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Exercise:

Problem:

[link] gives the number of suicides estimated in the U.S. for a recent year by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Race and Sex	**1–14**	**15–24**	**25–64**	**over 64**	**TOTALS**
white, male	210	3,360	13,610		22,050
white, female	80	580	3,380		4,930
black, male	10	460	1,060		1,670
black, female	0	40	270		330
all others					
TOTALS	310	4,650	18,780		29,760

Do not include "all others" for parts f and g.

a. Fill in the column for the suicides for individuals over age 64.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a white male.
d. Find the probability that a randomly selected individual was a black female.
e. Find the probability that a randomly selected individual was black
f. Find the probability that a randomly selected individual was a black or white male.
g. Out of the individuals over age 64, find the probability that a randomly selected individual was a black or white male.

Solution:

a.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

b.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

c. $\frac{22,050}{29,760}$

d. $\frac{330}{29,760}$

e. $\frac{2,000}{29,760}$

f. $\frac{23720}{(29760-780)} = \frac{23720}{28980}$

g. $\frac{5010}{(6020-100)} = \frac{5010}{5920}$

Use the following information to answer the next two exercises. The table of data obtained from *www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Exercise:

Problem: Find P(hit was made by Babe Ruth).

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Exercise:

Problem: Find P(hit was made by Ty Cobb|The hit was a Home Run).

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

Solution:

b

Exercise:

Problem: [link] identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43

Hair Type	Brown	Blond	Black	Red	Totals
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If *B* is the event of a child having brown hair, find the probability of the complement of *B*.
g. In words, what does the complement of *B* represent?

Exercise:

Problem:

In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data were compiled into the following table.

Shirt#	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

Solution:

a. $\frac{26}{106}$
b. $\frac{33}{106}$

c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right) + \left(\frac{33}{106}\right) - \left(\frac{21}{106}\right) = \left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

Glossary

contingency table
: the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

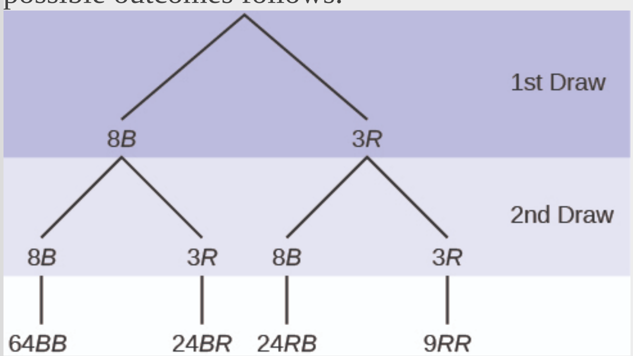
Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example:
In an urn, there are 11 balls. Three balls are red (*R*) and eight balls are blue (*B*). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.



Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as:
*R*1*R*1 *R*1*R*2 *R*1*R*3 *R*2*R*1 *R*2*R*2 *R*2*R*3 *R*3*R*1 *R*3*R*2 *R*3*R*3
The other outcomes are similar.
There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the **sample space**.

Exercise:

Problem: a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...

Solution:

a. $B1R1$ $B1R2$ $B1R3$ $B2R1$ $B2R2$ $B2R3$ $B3R1$ $B3R2$ $B3R3$ $B4R1$ $B4R2$ $B4R3$ $B5R1$ $B5R2$ $B5R3$ $B6R1$ $B6R2$ $B6R3$ $B7R1$ $B7R2$ $B7R3$ $B8R1$ $B8R2$ $B8R3$

Exercise:

Problem: b. Using the tree diagram, calculate $P(RR)$.

Solution:

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

Exercise:

Problem: c. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.

Solution:

c. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

Exercise:

Problem: d. Using the tree diagram, calculate $P(R$ on 1st draw AND B on 2nd draw$)$.

Solution:

d. $P(R$ on 1st draw AND B on 2nd draw$) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

Exercise:

Problem: e. Using the tree diagram, calculate $P(R$ on 2nd draw GIVEN B on 1st draw$)$.

Solution:

e. $P(R$ on 2nd draw GIVEN B on 1st draw$) = P(R$ on 2nd$|B$ on 1st$) = \frac{24}{88} = \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88} = \frac{3}{11}$.

Exercise:

Problem: f. Using the tree diagram, calculate $P(BB)$.

Solution:

f. $P(BB) = \frac{64}{121}$

Exercise:

Problem:

g. Using the tree diagram, calculate P(B on the 2nd draw given R on the first draw).

Solution:

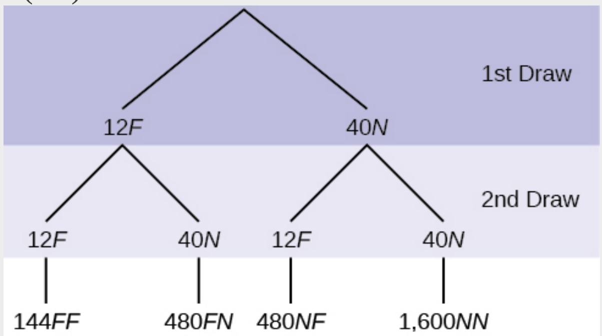
g. P(B on 2nd draw|R on 1st draw) = $\frac{8}{11}$

There are 9 + 24 outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Note:
Try It
Exercise:

Problem:

In a standard deck, there are 52 cards. 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(FF)$.



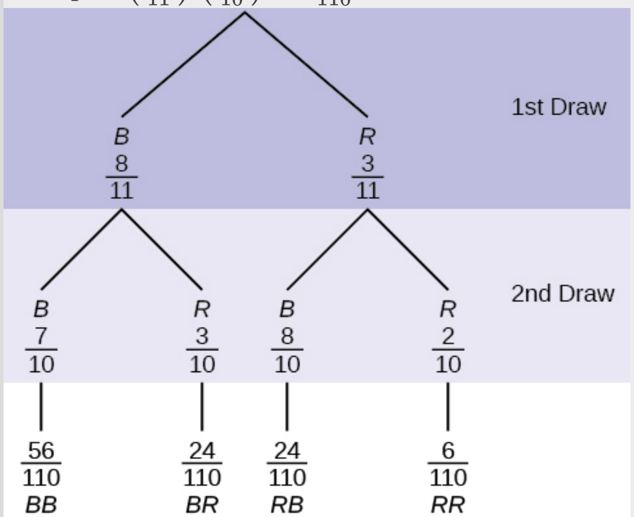
Solution:

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(FF) = \frac{144}{144 + 480 + 480 + 1{,}600} = \frac{144}{2{,}704} = \frac{9}{169}$$

Example:

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.



$$\text{Total} = \frac{56+24+24+6}{110} = \frac{110}{110} = 1$$

Note:

NOTE

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without replacement**, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

Exercise:

Problem: a. $P(RR)$ = _________

Solution:

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

Exercise:

Problem: b. Fill in the blanks:

$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (\_\_\_)(\_\_\_) = \frac{48}{110}$

Solution:

b. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

Exercise:

Problem: c. $P(R \text{ on 2nd} | B \text{ on 1st})$ =

Solution:

c. $P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{3}{10}$

Exercise:

Problem: d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\_\_\_)(\_\_\_) = \frac{24}{100}$

Solution:

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$

Exercise:

Problem: e. Find $P(BB)$.

Solution:

e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

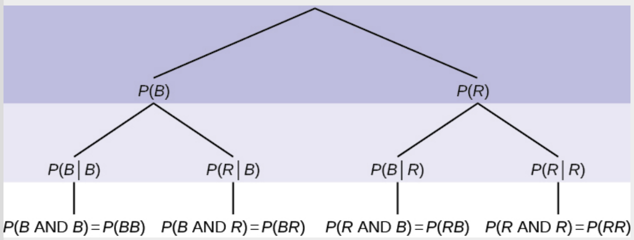
Exercise:

Problem: f. Find $P(B \text{ on 2nd}|R \text{ on 1st})$.

Solution:

f. Using the tree diagram, $P(B \text{ on 2nd}|R \text{ on 1st}) = P(R|B) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.

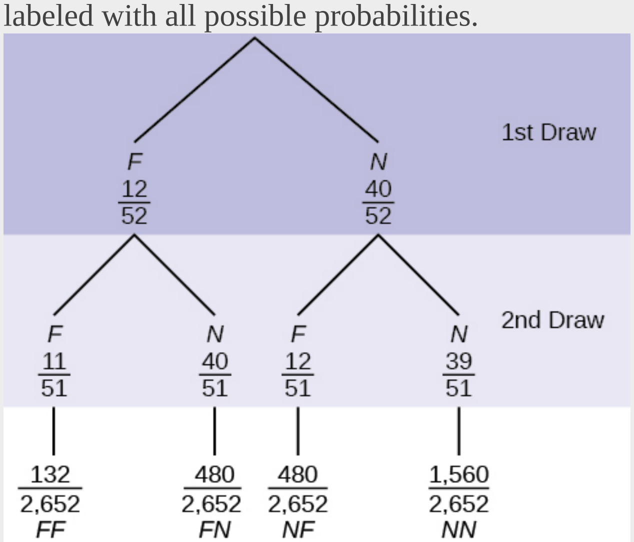


- $P(R|R)$ here means $P(R \text{ on 2nd}|R \text{ on 1st})$
- $P(B|R)$ here means $P(B \text{ on 2nd}|R \text{ on 1st})$
- $P(R|B)$ here means $P(R \text{ on 2nd}|B \text{ on 1st})$
- $P(B|B)$ here means $P(B \text{ on 2nd}|B \text{ on 1st})$

Note:
Try It
Exercise:

Problem:

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

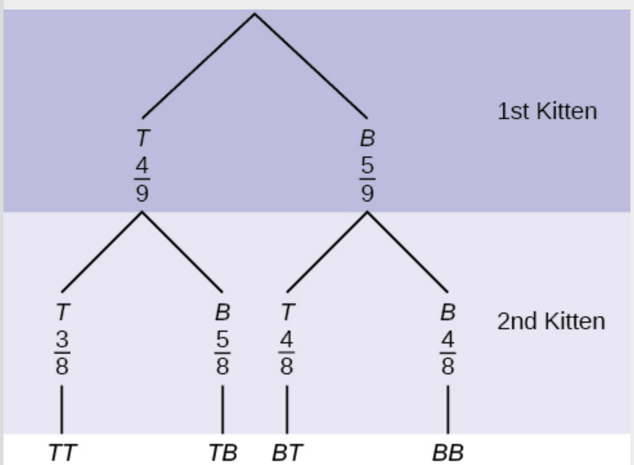


a. Find $P(FN \text{ OR } NF)$.
b. Find $P(N|F)$.
c. Find P(at most one face card).
Hint: "At most one face card" means zero or one face card.
d. Find P(at least on face card).
Hint: "At least one face card" means one or two face cards.

Solution:

a. $P(FN \text{ OR } NF) = \frac{480}{2,652} + \frac{480}{2,652} = \frac{960}{2,652} = \frac{80}{221}$
b. $P(N|F) = \frac{40}{51}$
c. $P(\text{at most one face card}) = \frac{(480 + 480 + 1,560)}{2,652} = \frac{2,520}{2,652}$
d. $P(\text{at least one face card}) = \frac{(132 + 480 + 480)}{2,652} = \frac{1,092}{2,652}$

Example:
A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



Exercise:

Problem:

a. What is the probability that both kittens are tabby?

a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
b. What is the probability that one kitten of each coloring is selected?

a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$

c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
d. What is the probability of choosing two kittens of the same color?

Solution:

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Note:
Try It
Exercise:

Problem:

Suppose there are four red balls and three yellow balls in a box. Two balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

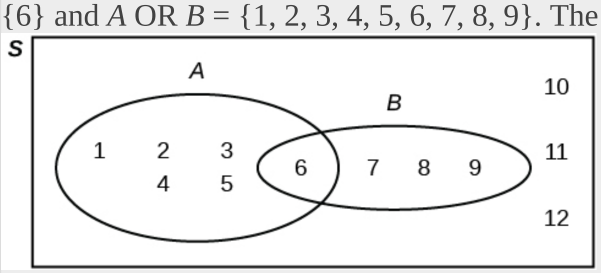
Solution:

$\left(\frac{4}{7}\right)\left(\frac{3}{6}\right)+\left(\frac{3}{7}\right)\left(\frac{4}{6}\right)$

Venn Diagram

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example:
Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event A = {1, 2, 3, 4, 5, 6} and event B = {6, 7, 8, 9}. Then A AND B = {6} and A OR B = {1, 2, 3, 4, 5, 6, 7, 8, 9}. The Venn diagram is as follows:

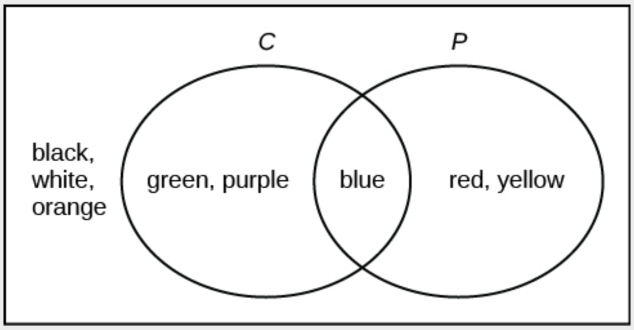


Note:
Try It
Exercise:

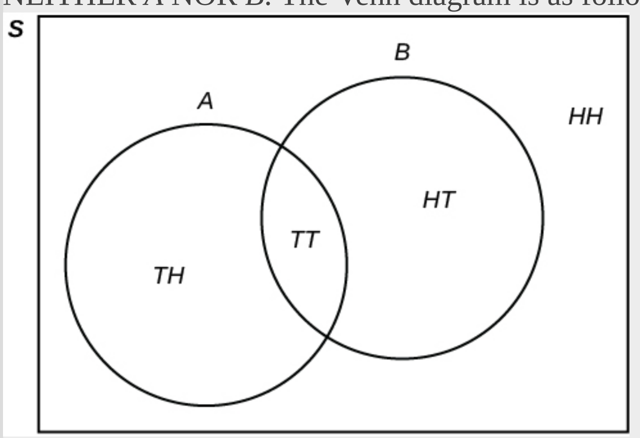
Problem:

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event *C* = {green, blue, purple} and event *P* = {red, yellow, blue}. Then *C* AND *P* = {blue} and *C* OR *P* = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

Solution:



Example:
Flip two fair coins. Let *A* = tails on the first coin. Let *B* = tails on the second coin. Then *A* = {*TT*, *TH*} and *B* = {*TT*, *HT*}. Therefore, *A* AND *B* = {*TT*}. *A* OR *B* = {*TH*, *TT*, *HT*}.
The sample space when you flip two fair coins is *X* = {*HH*, *HT*, *TH*, *TT*}. The outcome *HH* is in NEITHER *A* NOR *B*. The Venn diagram is as follows:

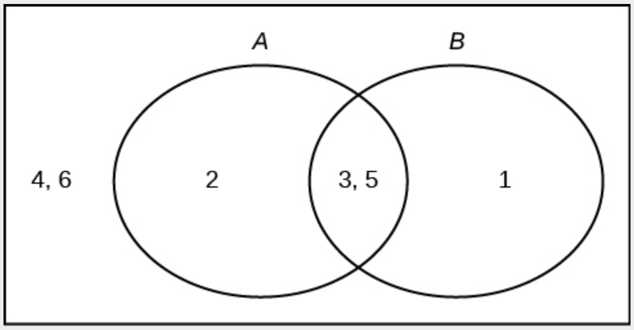


Note:
Try It
Exercise:

Problem:

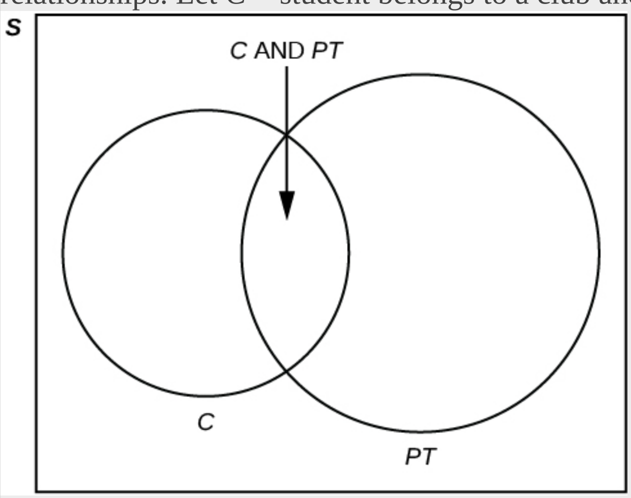
Roll a fair, six-sided die. Let A = a prime number of dots is rolled. Let B = an odd number of dots is rolled. Then A = {2, 3, 5} and B = {1, 3, 5}. Therefore, A AND B = {3, 5}. A OR B = {1, 2, 3, 5}. The sample space for rolling a fair die is S = {1, 2, 3, 4, 5, 6}. Draw a Venn diagram representing this situation.

Solution:



Example:

Forty percent of the students at a local college belong to a club and **50%** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let C = student belongs to a club and PT = student works part time.



If a student is selected at random, find

- the probability that the student belongs to a club. $P(C)$ = 0.40
- the probability that the student works part time. $P(PT)$ = 0.50
- the probability that the student belongs to a club AND works part time. $P(C \text{ AND } PT)$ = 0.05
- the probability that the student belongs to a club **given** that the student works part time. $P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{0.05}{0.50} = 0.1$

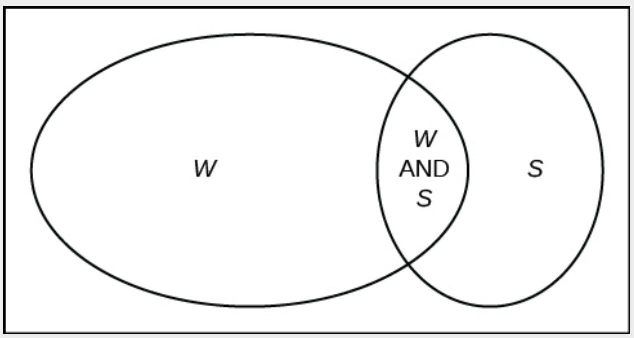
- the probability that the student belongs to a club **OR** works part time. $P(C \text{ OR } PT) = P(C) + P(PT) - P(C \text{ AND } PT) = 0.40 + 0.50 - 0.05 = 0.85$

Note:
Try It
Exercise:

Problem:

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let W = works a second job and S = spouse also works.

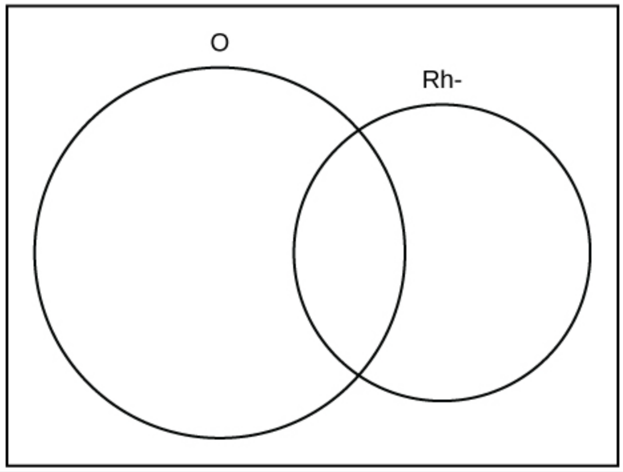
Solution:



Example:
Exercise:

Problem:

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5−10% of African Americans have the Rh- factor, and 51% have type O blood.



The “O” circle represents the African Americans with type O blood. The “Rh-“ oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let O = African American with Type O blood and R = African American with Rh- factor.

a. $P(O)$ = ____________
b. $P(R)$ = ____________
c. $P(O \text{ AND } R)$ = ____________
d. $P(O \text{ OR } R)$ = _____________
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Solution:

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

Note:
Try It
Exercise:

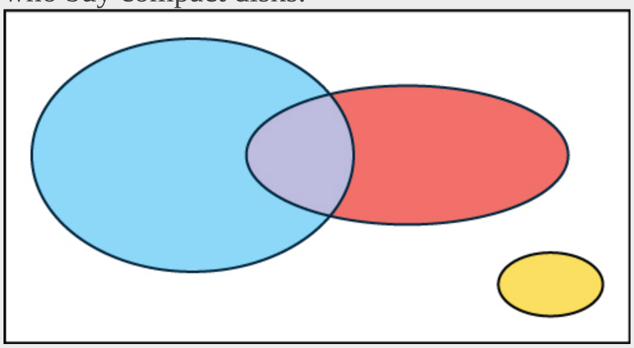
Problem:

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or anon-fiction book.
c. In the Venn diagram, describe the overlapping area using a complete sentence.
d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Solution:

a. and d. In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.



b. *P*(novel or non-fiction) = *P*(Blue OR Red) = *P*(Blue) + *P*(Red) - *P*(Blue AND Red) = 0.6 + 0.4 - 0.2 = 0.8.
c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

Chapter Review

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

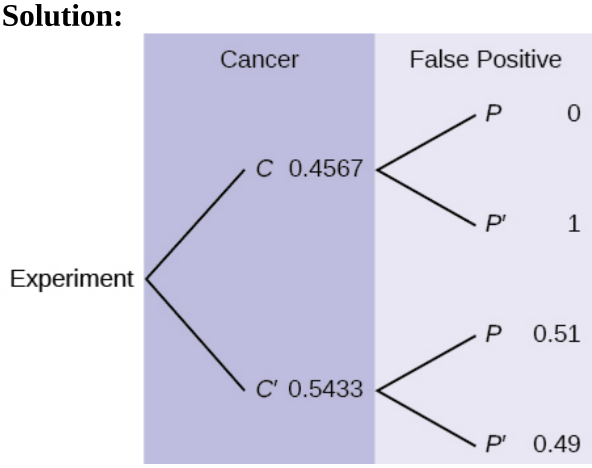
A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space *S* together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

Exercise:

Problem:

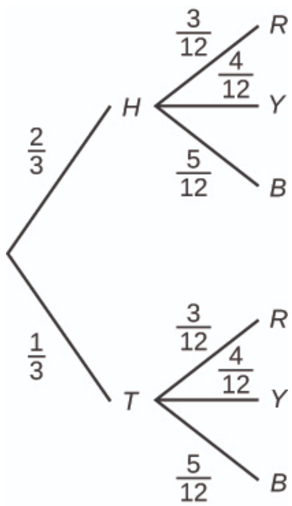
The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: *C* = a man develops cancer in his lifetime; *P* = man has at least one false positive. Construct a tree diagram of the situation.

Solution:



Homework

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (*R*), four yellow (*Y*) and five blue (*B*) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where *H* is heads and *T* is tails.



Exercise:

Problem: Find *P*(tossing a Head on the coin AND a Red bead)

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Exercise:

Problem: Find *P*(Blue bead).

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

Solution:

a

Exercise:

Problem:

A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.

b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find $P(S)$.
e. Let T be the event that the cookies selected were different flavors. Find $P(T)$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find $P(U)$.

Bringing It Together

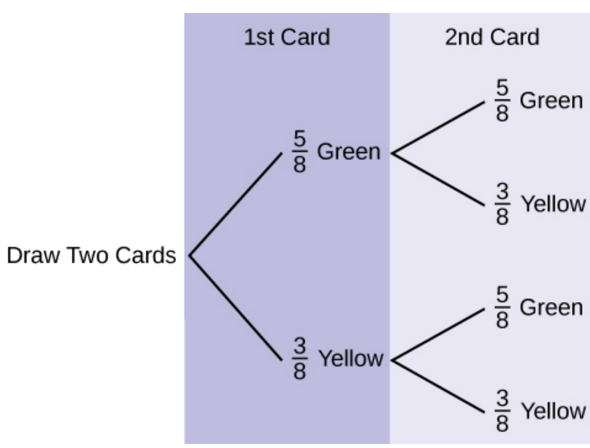
Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

Exercise:

Problem: Suppose that you randomly draw two cards, one at a time, **with replacement**.
Let G_1 = first card is green
Let G_2 = second card is green

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{ AND } G_2)$.
c. Find P(at least one green).
d. Find $P(G_2|G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Solution:

a. 

b. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
c. P(at least one green) = $P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
d. $P(G|G) = \frac{5}{8}$

e. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Exercise:

Problem: Suppose that you randomly draw two cards, one at a time, **without replacement**.
G_1 = first card is green
G_2 = second card is green

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{ AND } G_2)$.
c. Find P(at least one green).
d. Find $P(G_2|G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

Exercise:

Problem: Complete the following.

a. Construct a table or a tree diagram of the situation.
b. Find P(driver is female).
c. Find P(driver is age 65 or over|driver is female).
d. Find P(driver is age 65 or over AND female).
e. In words, explain the difference between the probabilities in part c and part d.
f. Find P(driver is age 65 or over).
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Solution:

a.

	<20	20–64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514

	<20	20–64	>64	Totals
Totals	0.0503	0.8140	0.1356	1

b. $P(F) = 0.486$
c. $P(>64|F) = 0.1361$
d. $P(>64 \text{ and } F) = P(F)\ P(>64|F) = (0.486)(0.1361) = 0.0661$
e. $P(>64|F)$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } F)$ is the percentage of drivers who are female and 65 or older.
f. $P(>64) = P(>64 \text{ and } F) + P(>64 \text{ and } M) = 0.1356$
g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } F) = 0.0661$.

Exercise:

Problem: Suppose that 10,000 U.S. licensed drivers are randomly selected.

a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

Exercise:

Problem:

Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Solution:

a.

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

b. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: P(Alone) = 0.7318 + 0.0390 = 0.7708.
c. Make the same assumptions as in (b) we have: (0.7708)(1,000) = 771
d. (0.1332)(1,000) = 133

Exercise:

Problem:

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

a. Based on the given data, find $P(H)$ and $P(T)$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

Exercise:

Problem:

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

* includes homosexual/bisexual IV drug users

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find *P*(Person is female).
b. Find *P*(Person has a risk factor heterosexual contact).
c. Find *P*(Person is female OR has a risk factor of IV drug user).
d. Find *P*(Person is female AND has a risk factor of homosexual/bisexual).
e. Find *P*(Person is male AND has a risk factor of IV drug user).
f. Find *P*(Person is female GIVEN person got the disease from heterosexual contact).
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

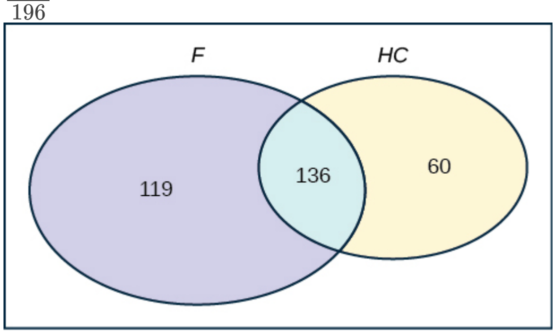
Solution:

The completed contingency table is as follows:

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

* includes homosexual/bisexual IV drug users

a. $\frac{255}{3059}$
b. $\frac{196}{3059}$

c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$
g. 

Exercise:

Problem:

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find P(Person is female).
b. Find P(Person obtained the disease through heterosexual contact).
c. Find P(Person is female GIVEN person got the disease from heterosexual contact)
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

Glossary

Tree Diagram
: the useful visual representation of a sample space and events in the form of a "tree" with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram
: the visual representation of a sample space and events in the form of circles or ovals showing their intersections

Introduction
class="introduction"

You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately.
- Recognize the geometric probability distribution and apply it appropriately.
- Recognize the hypergeometric probability distribution and apply it appropriately.
- Classify discrete word problems by their distributions.

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Random Variable Notation

Upper case letters such as *X* or *Y* denote a random variable. Lower case letters like *x* or *y* denote the value of a random variable. If ***X* is a random variable, then *X* is written in words, and *x* is given as a number.**

For example, let *X* = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*; *THH*; *HTH*; *HHT*; *HTT*; *THT*; *TTH*; *HHH*. Then, *x* = 0, 1, 2, 3. *X* is in words and *x* is a number. Notice that for this example, the *x* values are countable outcomes. Because you can count the possible values that *X* can take on and the outcomes are random (the *x* values 0, 1, 2, 3), *X* is a discrete random variable.

Note:
Collaborative Exercise
Toss a coin ten times and record the number of heads. After all members of the class have completed the experiment (tossed a coin ten times and counted the number of heads), fill in [link]. Let *X* = the number of heads in ten tosses of the coin.

x	**Frequency of *x***	**Relative Frequency of *x***

a. Which value(s) of *x* occurred most frequently?
b. If you tossed the coin 1,000 times, what values could *x* take on? Which value(s) of *x* do you think would occur most frequently?

c. What does the relative frequency column sum to?

Glossary

Random Variable (RV)
: a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters *X*, *Y*, *Z*,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters *x*, *y*, and *z*. For example, if *X* is the number of children in a family, then *x* represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if *X* = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value *x* the random variable *X* takes only after performing the experiment.

Probability Distribution Function (PDF) for a Discrete Random Variable

A discrete **probability distribution function** has two characteristics:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

Example:
A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, x = 0, 1, 2, 3, 4, 5.
$P(x)$ = probability that X takes on a value x.

x	$P(x)$
0	$P(x = 0) = \frac{2}{50}$
1	$P(x = 1) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$
3	$P(x = 3) = \frac{9}{50}$
4	$P(x = 4) = \frac{4}{50}$
5	$P(x = 5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because:

a. Each $P(x)$ is between zero and one, inclusive.
b. The sum of the probabilities is one, that is,

Equation:

$$\frac{2}{50}+\frac{11}{50}+\frac{23}{50}+\frac{9}{50}+\frac{4}{50}+\frac{1}{50}=1$$

Note:
Try It
Exercise:

Problem:

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, x = 0, 1, 2, 3, 4, 5. $P(x)$ = the probability that X takes on value x. Why is this a discrete probability distribution function (two reasons)?

X	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$

X	***P*(*x*)**
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Solution:

Each $P(x)$ is between 0 and 1, inclusive, and the sum of the probabilities is 1, that is: $\frac{4}{50} + \frac{8}{50} + \frac{16}{50} + \frac{14}{50} + \frac{6}{50} + \frac{2}{50} = 1$

Example:

Suppose Nancy has classes **three days** a week. She attends classes three days a week **80%** of the time, **two days 15%** of the time, **one day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

Exercise:

Problem:

a. Let *X* = the number of days Nancy ____________________.

Solution:

a. Let *X* = the number of days Nancy attends class per week.

Exercise:

Problem: b. *X* takes on what values?

Solution:

b. 0, 1, 2, and 3

Exercise:

Problem:

c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in [link]. The table should have two columns labeled *x* and *P*(*x*). What does the *P*(*x*) column sum to?

Solution:

c.

x	***P*(*x*)**
0	0.01
1	0.04
2	0.15
3	0.80

Note:
Try It
Exercise:

Problem:

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Solution:

X is the number of days Jeremiah attends basketball practice per week. X takes on the values 0, 1, and 2.

Chapter Review

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

Use the following information to answer the next five exercises: A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution.

Let X = the number of years a new hire will stay with the company.

Let $P(x)$ = the probability that a new hire will stay with the company x years.
Exercise:

Problem: Complete [link] using the data provided.

x	*P*(*x*)
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

Solution:

x	*P*(*x*)
0	0.12

x	*P*(*x*)
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

Exercise:

Problem: $P(x = 4) =$ ________

Exercise:

Problem: $P(x \geq 5) =$ ________

Solution:

0.10 + 0.05 = 0.15

Exercise:

Problem:

On average, how long would you expect a new hire to stay with the company?

Exercise:

Problem: What does the column "$P(x)$" sum to?

Solution:

1

Use the following information to answer the next six exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	***P*(*x*)**
1	0.15
2	0.35
3	0.40
4	0.10

Exercise:

Problem: Define the random variable X.

Exercise:

Problem:

What is the probability the baker will sell more than one batch? $P(x > 1)$ = ________

Solution:

$0.35 + 0.40 + 0.10 = 0.85$

Exercise:

Problem:

What is the probability the baker will sell exactly one batch? $P(x = 1) =$ ________

Exercise:

Problem: On average, how many batches should the baker make?

Solution:

$1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45$

Use the following information to answer the next four exercises: Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

Exercise:

Problem: Define the random variable X.

Exercise:

Problem: Construct a probability distribution table for the data.

Solution:

x	*P*(*x*)
0	0.03
1	0.04
2	0.08
3	0.85

Exercise:

Problem:

We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Use the following information to answer the next five exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

Exercise:

Problem: Define the random variable *X*.

Solution:

Let *X* = the number of events Javier volunteers for each month.

Exercise:

Problem: What values does *x* take on?

Exercise:

Problem: Construct a PDF table.

Solution:

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

Exercise:

Problem:

Find the probability that Javier volunteers for less than three events each month. $P(x < 3)$ = ________

Exercise:

Problem:

Find the probability that Javier volunteers for at least one event each month. $P(x > 0)$ = ________

Solution:

1 – 0.05 = 0.95

Homework

Exercise:

Problem:

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in [link].

x	***P*(*x*)**
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

a. In words, define the random variable *X*.
b. What does it mean that the values zero, one, and two are not included for *x* in the PDF?

Glossary

Probability Distribution Function (PDF)

a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Mean or Expected Value and Standard Deviation

The **expected value** is often referred to as the **"long-term" average or mean**. This means that over the long term of doing an experiment over and over, you would **expect** this average.

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. As you learned in [link], probability does not describe the short-term results of an experiment. It gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. **In his experiment, Pearson illustrated the Law of Large Numbers**.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero **(the theoretical probability and the relative frequency get closer and closer together)**. When evaluating the long-term results of statistical experiments, we often want to know the “average” outcome. This “long-term average” is known as the **mean** or **expected value** of the experiment and is denoted by the Greek letter μ. In other words, after conducting many trials of an experiment, you would expect this average value.

Note:
NOTE
To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

Example:
A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play

one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the men's soccer team plays soccer.
To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table adding a column $x*P(x)$. In this column, you will multiply each x value by its probability.

x	***P*(*x*)**	***x**P*(*x*)**
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Expected Value TableThis table is called an expected value table. The table helps you calculate the expected value or long-term average.

Add the last column $x*P(x)$ to find the long term average or expected value: (0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1.
The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu = 1.1$.

Example:
Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

x	$P(x)$	$x*P(x)$	$(x - \mu)^2 \cdot P(x)$
0	$P(x = 0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$
1	$P(x = 1) = \left(\frac{11}{50}\right)$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x = 2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x = 3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x = 4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x = 5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

Add the values in the third column of the table to find the expected value of X:
μ = Expected Value = $\frac{105}{50} = 2.1$
Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x, multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$).
Add the values in the fourth column of the table:
$0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05$
The standard deviation of X is the square root of this sum: $\sigma = \sqrt{1.05} \approx 1.0247$

Note:
Try It
Exercise:

Problem:

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	***P*(*x*)**
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Solution:

The expected value is 2.32

$(0)\frac{4}{50} + (1)\frac{8}{50} + (2)\frac{16}{50} + (3)\frac{14}{50} + (4)\frac{6}{50} + (5)\frac{2}{50} = 0 + \frac{8}{50} + \frac{32}{50} + \frac{42}{50} + \frac{24}{50} + \frac{10}{50} = \frac{116}{50} = 2.32$

Example:

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and −2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

Equation:

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = (1)(10^{-5}) = 0.00001.$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

Equation:

$$1 - 0.00001 = 0.99999.$$

The expected value table is as follows:

	x	*P*(*x*)	*x***P*(*x*)
Loss	–2	0.99999	(–2)(0.99999) = –1.99998
Profit	100,000	0.00001	(100000)(0.00001) = 1

Add the last column. –1.99998 + 1 = –0.99998

Since –0.99998 is about –1, you would, on average, expect to lose approximately $1 for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected LOSS per game after playing this game over and over.

Note:
Try It
Exercise:

Problem:

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit of playing the game over the long term?

Solution:

Let *X* = the amount of money you profit. The *x*-values are –$1 and $256.

The probability of guessing the right suit each time is $\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{256}$ = 0.0039

The probability of losing is 1 – $\frac{1}{256}$ $=$ $\frac{255}{256}$ = 0.9961

(0.0039)256 + (0.9961)(–1) = 0.9984 + (–0.9961) = 0.0023 or 0.23 cents.

Example:
Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads}) = \frac{2}{3}$ and $P(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

Exercise:

Problem: a. Define a random variable *X*.

Solution:

a. *X* = amount of profit

Exercise:

Problem: b. Complete the following expected value table.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

Solution:

b.

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

Exercise:

Problem: c. What is the expected value, μ? Do you come out ahead?

Solution:

c. Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Note:
Try It
Exercise:

Problem:

Suppose you play a game with a spinner. You play each game by spinning the spinner once. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. If you land on red, you pay $10. If you land on blue, you don't pay or win anything. If you land on green, you win $10. Complete the following expected value table.

	x	***P*(*x*)**	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Solution:

	x	***P*(*x*)**	***x*P*(*x*)**
Red	–10	$\frac{2}{5}$	$-\frac{20}{5}$
Blue	0	$\frac{2}{5}$	$\frac{0}{5}$

	x	$P(x)$	$x*P(x)$
Green	10	$\frac{1}{5}$	$\frac{10}{5}$

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root. To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x – \mu)^2P(x)$ and take the square root.

x	$P(x)$	$x*P(x)$	$(x – \mu)^2P(x)$
0	0.2	(0)(0.2) = 0	$(0 – 1.1)^2(0.2) = 0.242$
1	0.5	(1)(0.5) = 0.5	$(1 – 1.1)^2(0.5) = 0.005$
2	0.3	(2)(0.3) = 0.6	$(2 – 1.1)^2(0.3) = 0.243$

Add the last column in the table. 0.242 + 0.005 + 0.243 = 0.490. The standard deviation is the square root of 0.49, or $\sigma = \sqrt{0.49} = 0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas for calculating μ and σ.

Example:
Exercise:

Problem:

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like [link] and calculate the mean μ and standard deviation σ of X.

Solution:

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Use the sample space to complete the following table:

x	$P(x)$	$xP(x)$	$(x - \mu)^2 \cdot P(x)$
0	$\frac{9}{36}$	0	$(0 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Calculating μ and σ.

Add the values in the third column to find the expected value: $\mu = \frac{36}{36} = 1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum: $\sigma = \sqrt{\frac{18}{36}} \approx 0.7071$.

Example:
Exercise:

Problem:

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win $50. If you lose the bet, you pay $20. Let *X* = the amount of profit from a bet.

P(win) = *P*(one moderate earthquake will occur) = 21.42%

P(loss) = *P*(one moderate earthquake will *not* occur) = 100% – 21.42%

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of *X*? Construct a table similar to [link] and [link] to help you answer these questions.

Solution:

	x	***P(x)***	***x(Px)***	**$(x - \mu)^2 P(x)$**
win	50	0.2142	10.71	[50 – (–5.006)]^2(0.2142) = 648.0964
loss	–20	0.7858	–15.716	[–20 – (–5.006)]^2(0.7858) = 176.6636

Mean = Expected Value = 10.71 + (–15.716) = –5.006.

If you make this bet many times under the same conditions, your long term outcome will be an average *loss* of $5.01 per bet.

$\text{Standard Deviation} = \sqrt{648.0964 + 176.6636} \approx 28.7186$

Note:
Try It
Exercise:

Problem:

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. As in [link], you bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win $100. If you lose the bet, you pay $10. Let *X* = the amount of profit from a bet. Find the mean and standard deviation of *X*.

Solution:

	x	$P(x)$	$x \cdot (Px)$	$(x - \mu)^2 \cdot P(x)$
win	100	0.0108	1.08	$[100 – (–8.812)]^2 \cdot 0.0108 = 127.8726$
loss	–10	0.9892	–9.892	$[–10 – (–8.812)]^2 \cdot 0.9892 = 1.3961$

Mean = Expected Value = μ = 1.08 + (–9.892) = –8.812

If you make this bet many times under the same conditions, your long term outcome will be an average loss of $8.81 per bet.

$$\text{Standard Deviation} = \sqrt{127.7826 + 1.3961} \approx 11.3696$$

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not

cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

References

Class Catalogue at the Florida State University. Available online at https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy (accessed May 15, 2013).

"World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

Chapter Review

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Formula Review

Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$

Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x-\mu)^2 P(x)}$

Exercise:

Problem: Complete the expected value table.

x	***P*(*x*)**	***x*P*(*x*)**
0	0.2	
1	0.2	
2	0.4	
3	0.2	

Exercise:

Problem: Find the expected value from the expected value table.

x	***P*(*x*)**	***x*P*(*x*)**
2	0.1	2(0.1) = 0.2
4	0.3	4(0.3) = 1.2
6	0.4	6(0.4) = 2.4
8	0.2	8(0.2) = 1.6

Solution:

0.2 + 1.2 + 2.4 + 1.6 = 5.4

Exercise:

Problem: Find the standard deviation.

x	$P(x)$	$x*P(x)$	$(x-\mu)^2P(x)$
2	0.1	2(0.1) = 0.2	$(2–5.4)^2(0.1) = 1.156$
4	0.3	4(0.3) = 1.2	$(4–5.4)^2(0.3) = 0.588$
6	0.4	6(0.4) = 2.4	$(6–5.4)^2(0.4) = 0.144$
8	0.2	8(0.2) = 1.6	$(8–5.4)^2(0.2) = 1.352$

Exercise:

Problem: Identify the mistake in the probability distribution table.

x	$P(x)$	$x*P(x)$

x	***P*(*x*)**	***x**P*(*x*)**
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

Solution:

The values of $P(x)$ do not sum to one.

Exercise:

Problem: Identify the mistake in the probability distribution table.

x	***P*(*x*)**	***x**P*(*x*)**
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85

x	*P*(*x*)	*x***P*(*x*)
5	0.15	1

Use the following information to answer the next five exercises: A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

x	*P*(*x*)	*x***P*(*x*)
1	0.35	
2	0.20	
3	0.15	
4		
5	0.10	
6	0.05	

Exercise:

Problem: Define the random variable *X*.

Solution:

Let X = the number of years a physics major will spend doing post-graduate research.

Exercise:

Problem: Define $P(x)$, or the probability of x.

Exercise:

Problem:

Find the probability that a physics major will do post-graduate research for four years. $P(x = 4)$ = ________

Solution:

1 – 0.35 – 0.20 – 0.15 – 0.10 – 0.05 = 0.15

Exercise:

Problem:

FInd the probability that a physics major will do post-graduate research for at most three years. $P(x \leq 3)$ = ________

Exercise:

Problem:

On average, how many years would you expect a physics major to spend doing post-graduate research?

Solution:

1(0.35) + 2(0.20) + 3(0.15) + 4(0.15) + 5(0.10) + 6(0.05) = 0.35 + 0.40 + 0.45 + 0.60 + 0.50 + 0.30 = 2.6 years

Use the following information to answer the next seven exercises: A ballet instructor is interested in knowing what percent of each year's class will

continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let X = the number of years a student will study ballet with the teacher.
- Let $P(x)$ = the probability that a student will study ballet x years.

Exercise:

Problem: Complete [link] using the data provided.

x	$P(x)$	$x*P(x)$
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

Exercise:

Problem: In words, define the random variable X.

Solution:

X is the number of years a student studies ballet with the teacher.

Exercise:

Problem: $P(x = 4) =$ ________

Exercise:

Problem: $P(x < 4) =$ ________

Solution:

0.10 + 0.05 + 0.10 = 0.25

Exercise:

Problem:

On average, how many years would you expect a child to study ballet with this teacher?

Exercise:

Problem: What does the column "$P(x)$" sum to and why?

Solution:

The sum of the probabilities sum to one because it is a probability distribution.

Exercise:

Problem: What does the column "$x*P(x)$" sum to and why?

Exercise:

Problem:

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?

Solution:

$-2\left(\frac{40}{52}\right)+30\left(\frac{12}{52}\right)=-1.54+6.92=5.38$

Exercise:

Problem:

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. Should you play the game?

HOMEWORK

Exercise:

Problem:

A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of $150.

a. What are you interested in here?
b. In words, define the random variable *X*.
c. List the values that *X* may take on.
d. Construct a PDF.
e. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

Exercise:

Problem:

A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win $6
- If the card is a face card, and the coin lands on Tails, you win $2
- If the card is not a face card, you lose $2, no matter what the coin shows.

a. Find the expected value for this game (expected net gain or loss).
b. Explain what your calculations indicate about your long-term average profits and losses on this game.
c. Should you play this game to win money?

Solution:

The variable of interest is X, or the gain or loss, in dollars.

The face cards jack, queen, and king. There are (3)(4) = 12 face cards and 52 – 12 = 40 cards that are not face cards.

We first need to construct the probability distribution for X. We use the card and coin events to determine the probability for each outcome, but we use the monetary value of X to determine the expected value.

Card Event	***X* net gain/loss**	***P*(*X*)**
Face Card and Heads	6	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$

Card Event	*X* net gain/loss	*P*(*X*)
Face Card and Tails	2	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
(Not Face Card) and (H or T)	–2	$\left(\frac{40}{52}\right)(1)=\left(\frac{40}{52}\right)$

- Expected value $=(6)\left(\frac{6}{52}\right)+(2)\left(\frac{6}{52}\right)+(-2)\left(\frac{40}{52}\right)=-\frac{32}{52}$
- Expected value = –\$0.62, rounded to the nearest cent
- If you play this game repeatedly, over a long string of games, you would expect to lose 62 cents per game, on average.
- You should not play this game to win money because the expected value indicates an expected average loss.

Exercise:

Problem:

You buy a lottery ticket to a lottery that costs \$10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one \$500 prize, two \$100 prizes, and four \$25 prizes. Find your expected gain or loss.

Exercise:

Problem: Complete the PDF and answer the questions.

x	*P*(*x*)	*xP*(*x*)
0	0.3	

x	***P*(*x*)**	***xP*(*x*)**
1	0.2	
2		
3	0.4	

a. Find the probability that $x = 2$.
b. Find the expected value.

Solution:

a. 0.1
b. 1.6

Exercise:

Problem:

Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.

a. What are you ultimately interested in here (the value of the roll or the money you win)?
b. In words, define the Random Variable *X*.
c. List the values that *X* may take on.
d. Construct a PDF.
e. Over the long run of playing this game, what are your expected average winnings per game?
f. Based on numerical values, should you take the deal? Explain your decision in complete sentences.

Exercise:

Problem:

A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

a. Construct a PDF for each investment.
b. Find the expected value for each investment.
c. Which is the safest investment? Why do you think so?
d. Which is the riskiest investment? Why do you think so?
e. Which investment has the highest expected return, on average?

Solution:

a.

Software Company	
x	$P(x)$
5,000,000	0.10
1,000,000	0.30
–1,000,000	0.60

Hardware Company	
x	***P*(*x*)**
3,000,000	0.20
1,000,000	0.40
–1,000,00	0.40

Biotech Firm	
x	***P*(*x*)**
6,00,000	0.10
0	0.70
–1,000,000	0.20

b. $200,000; $600,000; $400,000
c. third investment because it has the lowest probability of loss
d. first investment because it has the highest probability of loss
e. second investment

Exercise:

Problem:

Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children married people have.

x	$P(x)$	$xP(x)$
0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

a. Find the probability that a married adult has three children.
b. In words, what does the expected value in this example represent?
c. Find the expected value.
d. Is it more likely that a married adult will have two to three children or four to six children? How do you know?

Exercise:

Problem:

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given as in [link].

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

On average, how many years do you expect it to take for an individual to earn a B.S.?

Solution:

4.85 years

Exercise:

Problem:

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	***P*(*x*)**
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

a. Describe the random variable *X* in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.
Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	***P*(*x*)**
0	0.35
1	0.25
2	0.20

x	*P*(*x*)
3	0.10
4	0.05
5	0.05

e. At which store is the expected number of DVDs rented per customer higher?
f. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
g. If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
h. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

Exercise:

Problem:

A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.

b. No, I expect to come out behind in money.
c. It doesn't matter. I expect to break even.

Solution:

b

Exercise:

Problem:

Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.

a. What is the average class size assuming each class is filled to capacity?
b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable *X* equal the size of the student's class. Define the PDF for *X*.
c. Find the mean of *X*.
d. Find the standard deviation of *X*.

Exercise:

Problem:

In a lottery, there are 250 prizes of $5, 50 prizes of $25, and ten prizes of $100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

Solution:

Let *X* = the amount of money to be won on a ticket. The following table shows the PDF for *X*.

x	$P(x)$
0	0.969
5	$\frac{250}{10,000}$ = 0.025
25	$\frac{50}{10,000}$ = 0.005
100	$\frac{10}{10,000}$ = 0.001

Calculate the expected value of X.

0(0.969) + 5(0.025) + 25(0.005) + 100(0.001) = 0.35

A fair price for a ticket is $0.35. Any price over $0.35 will enable the lottery to raise money.

Glossary

Expected Value
: expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP(x)$.

Mean
: a number that measures the central tendency; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (detonated by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$ and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Mean of a Probability Distribution
: the long-term average of many trials of a statistical experiment

Standard Deviation of a Probability Distribution

a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

The Law of Large Numbers

As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

Binomial Distribution

There are three characteristics of a binomial experiment.

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X = the number of successes obtained in the n independent trials.

The mean, μ, and variance, σ^2, for the binomial probability distribution are $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ, is then $\sigma = \sqrt{npq}$.

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example:

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable X = the number of students who withdraw from the randomly selected elementary physics class.

Note:
Try It
Exercise:

Problem:

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a "success" be in this case?

Solution:

a school that offers fruit in their lunch every day

Example:
Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is p = 0.55. The probability of a failure is q = 0.45. The number of trials is n = 20. The probability question can be stated mathematically as $P(x = 15)$.

Note:
Try It
Exercise:

Problem:

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

Solution:

$P(x = 12)$

Example:
Exercise:

Problem:

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically.

Solution:

$P(x > 10)$

Note:
Try It
Exercise:

Problem:

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

Solution:

$P(x > 3)$

Example:
Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

Exercise:

Problem:

a. This is a binomial problem because there is only a success or a ___________, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.

Solution:

a. failure

Exercise:

Problem:

b. If we are interested in the number of students who do their homework on time, then how do we define *X*?

Solution:

b. X = the number of statistics students who do their homework on time

Exercise:

Problem: c. What values does x take on?

Solution:

c. 0, 1, 2, …, 50

Exercise:

Problem: d. What is a "failure," in words?

Solution:

d. Failure is defined as a student who does not complete his or her homework on time.

The probability of a success is p = 0.70. The number of trials is n = 50.

Exercise:

Problem: e. If $p + q = 1$, then what is q?

Solution:

e. q = 0.30

Exercise:

Problem:

f. The words "at least" translate as what kind of inequality for the probability question $P(x$ _____ $40)$.

Solution:

f. greater than or equal to (≥)
The probability question is $P(x \geq 40)$.

Note:
Try It
Exercise:

Problem:

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

Solution:

This is a binomial problem because there is only a success or a failure, and there are a definite number of trials. The probability of a success stays the same for each trial.

Notation for the Binomial: *B* = Binomial Probability Distribution Function

$X \sim B(n, p)$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p; n = number of trials, p = probability of a success on

each trial.

Example:

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20$, $p = 0.41$, and $q = 1 - 0.41 = 0.59$. $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (calculator or computer)

Note:

Go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate (x = value): binompdf(n, p, number) if "number" is left out, the result is the binomial probability table.

To calculate $P(x \leq$ value): binomcdf(n, p, number) if "number" is left out, the result is the cumulative binomial probability table.

For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,0.41,12). The result is $P(x \leq 12) = 0.9738$.

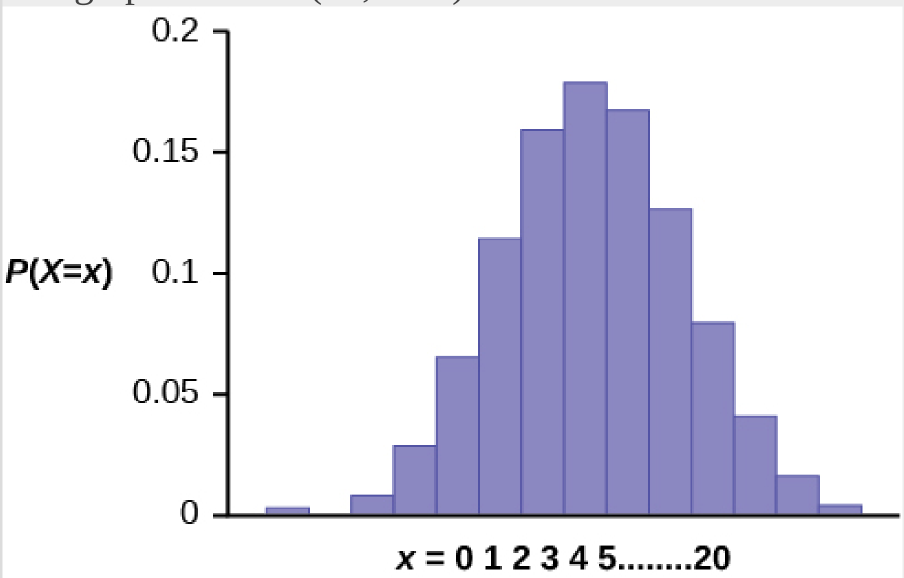
Note:

NOTE

If you want to find $P(x = 12)$, use the pdf (binompdf). If you want to find $P(x > 12)$, use 1 - binomcdf(20,0.41,12).

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

The graph of $X \sim B(20, 0.41)$ is as follows:



The y-axis contains the probability of x, where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$.

$\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20$.

Note:

Try It

Exercise:

Problem:

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

Solution:

$P(x \leq 14) = 0.9695$

Example:
Exercise:

Problem:

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

a. What values does x take on?
b. What is the probability distribution? Find the following probabilities:
   i. the probability that two pages feature signature artists
   ii. the probability that at most six pages feature signature artists
   iii. the probability that more than three pages feature signature artists.
c. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Solution:

a. $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
b. $X \sim B\left(100, \frac{8}{560}\right)$
   i. $P(x = 2) = \text{binompdf}\left(100, \frac{8}{560}, 2\right) = 0.2466$
   ii. $P(x \leq 6) = \text{binomcdf}\left(100, \frac{8}{560}, 6\right) = 0.9994$

iii. $P(x > 3) = 1 – P(x \leq 3) = 1 – \text{binomcdf}\left(100, \frac{8}{560}, 3\right) = 1 – 0.9443 = 0.0557$

c. i. Mean = $np = (100)\left(\frac{8}{560}\right) = \frac{800}{560} \approx 1.4286$
ii. Standard Deviation = $\sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

Note:
Try It
Exercise:

Problem:

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

a. What is the probability distribution for X?
b. Use your calculator to find the following probabilities:

   i. the probability that 25 adults in the sample prefer saving over spending
   ii. the probability that at most 20 adults prefer saving
   iii. the probability that more than 30 adults prefer saving

c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

Solution:

a. $X \sim B(50, 0.6)$
b. Using the TI-83, 83+, 84 calculator with instructions as provided in [link]:

i. $P(x = 25)$ = binompdf(50, 0.6, 25) = 0.0405
ii. $P(x \leq 20)$ = binomcdf(50, 0.6, 20) = 0.0034
iii. $P(x > 30)$ = 1 - binomcdf(50, 0.6, 30) = 1 – 0.5535 = 0.4465

c. i. Mean = np = 50(0.6) = 30
ii. Standard Deviation = $\sqrt{npq} = \sqrt{50\,(0.6)\,(0.4)} \approx 3.4641$

Example:
The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.
Exercise:

Problem:

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that at most eight people develop pancreatic cancer
d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Solution:

a. $X \sim B(200, 0.0128)$

b. i. Mean = np = 200(0.0128) = 2.56
ii. Standard Deviation = $\sqrt{npq} = \sqrt{(200)(0.0128)(0.9872)} \approx 1.5897$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in [link]:
$P(x \leq 8)$ = binomcdf(200, 0.0128, 8) = 0.9988

d. $P(x = 5)$ = binompdf(200, 0.0128, 5) = 0.0707
$P(x = 6)$ = binompdf(200, 0.0128, 6) = 0.0298
So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

Note:
Try It
Exercise:

Problem:

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
d. Find the probability that DeAndre scored with more than 50 of these shots.

Solution:

a. $X \sim B(80, 0.613)$

b. i. Mean = np = 80(0.613) = 49.04
ii. Standard Deviation =
$$\sqrt{npq} = \sqrt{80(0.613)(0.387)} \approx 4.3564$$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in [link]:
$P(x = 60)$ = binompdf(80, 0.613, 60) = 0.0036
d. $P(x > 50) = 1 – P(x \le 50)$ = 1 – binomcdf(80, 0.613, 50) = 1 – 0.6282 = 0.3718

Example:
The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw selects a student. The probability is $\frac{6}{15}$, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Note:
Try It
Exercise:

Problem:

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

Solution:

This is not binomial because the names are not replaced, which means the probability changes for each time a name is drawn. This violates the condition of independence.

References

"Access to electricity (% of population)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

"Distance Education." Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

"NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," GALLUP® Economy, 2013. Available online at http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los

Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at https://www.cia.gov/library/publications/the-world-factbook/geos/af.html (accessed May 15, 2013).

“What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics (accessed May 15, 2013).

Chapter Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, *n*.
2. There are only two possible outcomes, called "success" and, "failure" for each trial. The letter *p* denotes the probability of a success on one trial and *q* denotes the probability of a failure on one trial.
3. The *n* trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable *X* = the number of successes obtained in the *n* independent trials. The mean of *X* can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

Formula Review

$X \sim B(n, p)$ means that the discrete random variable *X* has a binomial probability distribution with *n* trials and probability of success *p*.

X = the number of successes in *n* independent trials

n = the number of independent trials

X takes on the values $x = 0, 1, 2, 3, ..., n$

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1$

$q = 1 – p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

Use the following information to answer the next eight exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

Exercise:

Problem: In words, define the random variable X.

Solution:

X = the number that reply "yes"

Exercise:

Problem: X ~ _____(_____,_____)

Exercise:

Problem: What values does the random variable X take on?

Solution:

0, 1, 2, 3, 4, 5, 6, 7, 8

Exercise:

Problem: Construct the probability distribution function (PDF).

x	***P*(*x*)**

Exercise:

Problem: On average (μ), how many would you expect to answer yes?

Solution:

5.7

Exercise:

Problem: What is the standard deviation (σ)?

Exercise:

Problem:

What is the probability that at most five of the freshmen reply “yes”?

Solution:

0.4151

Exercise:

Problem:

What is the probability that at least two of the freshmen reply “yes”?

HOMEWORK

Exercise:

Problem:

According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery.

Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

Exercise:

Problem: Define the random variable and list its possible values.

Solution:

X = the number of patients calling in claiming to have the flu, who actually have the flu.

$X = 0, 1, 2, ...25$

Exercise:

Problem: State the distribution of X.

Exercise:

Problem:

Find the probability that at least four of the 25 patients actually have the flu.

Solution:

0.0165

Exercise:

Problem:

On average, for every 25 patients calling in, how many do you expect to have the flu?

Exercise:

Problem:

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given [link]. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	*P*(*x*)
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

a. Describe the random variable *X* in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.

Solution:

a. *X* = the number of DVDs a Video to Go customer rents
b. 0.12
c. 0.11
d. 0.77

Exercise:

Problem:

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many of the 12 students do we expect to attend the festivities?
e. Find the probability that at most four students will attend.
f. Find the probability that more than two students will attend.

Use the following information to answer the next two exercises: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.

Exercise:

Problem: The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{382}{1043}$
d. 4.43

Solution:

d. 4.43

Let *X* = the number of games won in that upcoming month.

Exercise:

Problem:

What is the probability that the San Jose Sharks win six games in that upcoming month?

a. 0.1476
b. 0.2336

c. 0.7664
d. 0.8903

Exercise:

Problem:

What is the probability that the San Jose Sharks win at least five games in that upcoming month

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

Solution:

c

Exercise:

Problem:

A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

Exercise:

Problem:

A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

Solution:

- X = number of questions answered correctly

- $X \sim B\left(32, \frac{1}{3}\right)$
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event "more than 24" is the complement of "less than or equal to 24."
- Using your calculator's distribution menu: 1 – binomcdf $\left(32, \frac{1}{3}, 24\right)$
- $P(x > 24) = 0$
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

Exercise:

Problem:

Six different colored dice are rolled. Of interest is the number of dice that show a one.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. On average, how many dice would you expect to show a one?
e. Find the probability that all six dice show a one.
f. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

Exercise:

Problem:

More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)

d. On average, how many schools would you expect to offer such courses?
e. Find the probability that at most ten offer such courses.
f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

Solution:

a. *X* = the number of college and universities that offer online offerings.
b. 0, 1, 2, …, 13
c. *X* ~ *B*(13, 0.96)
d. 12.48
e. 0.0135
f. *P*(*x* = 12) = 0.3186 *P*(*x* = 13) = 0.5882 More likely to get 13.

Exercise:

Problem:

Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many are expected to attend their graduation?
e. Find the probability that 17 or 18 attend.
f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

Exercise:

Problem:

At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many are expected to **not** to use the foil as their main weapon?
e. Find the probability that six do **not** use the foil as their main weapon.
f. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

Solution:

a. *X* = the number of fencers who do **not** use the foil as their main weapon
b. 0, 1, 2, 3,... 25
c. $X \sim B(25,0.40)$
d. 10
e. 0.0442
f. The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

Exercise:

Problem:

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. $X \sim$ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Exercise:

Problem:

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. $X \sim$ _____(_____,_____)
d. How many audits are expected in a 20-year period?
e. Find the probability that a person is not audited at all.
f. Find the probability that a person is audited more than twice.

Solution:

a. X = the number of audits in a 20-year period
b. 0, 1, 2, …, 20
c. $X \sim B(20, 0.02)$
d. 0.4
e. 0.6676
f. 0.0071

Exercise:

Problem:

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. What is the probability that at least eight have adequate earthquake supplies?
e. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
f. How many residents do you expect will have adequate earthquake supplies?

Exercise:

Problem:

There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let *X* = number of matches and *Y* = profit per game.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. List the values that *Y* may take on. Then, construct one PDF table that includes both *X* and *Y* and their probabilities.
e. Calculate the average expected matches over the long run of playing this game for the player.
f. Calculate the average expected earnings over the long run of playing this game for the player.
g. Determine who has the advantage, the player or the house.

Solution:

1. *X* = the number of matches
2. 0, 1, 2, 3
3. $X \sim B\left(3, \frac{1}{6}\right)$
4. In dollars: −1, 1, 2, 3
5. $\frac{1}{2}$
6. Multiply each *Y* value by the corresponding *X* probability from the PDF table. The answer is −0.0787. You lose about eight cents, on average, per game.
7. The house has the advantage.

Exercise:

Problem:

According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let *X* = the number of people who have access to electricity.

a. What is the probability distribution for *X*?
b. Using the formulas, calculate the mean and standard deviation of *X*.

c. Use your calculator to find the probability that 15 people in the sample have access to electricity.
d. Find the probability that at most ten people in the sample have access to electricity.
e. Find the probability that more than 25 people in the sample have access to electricity.

Exercise:

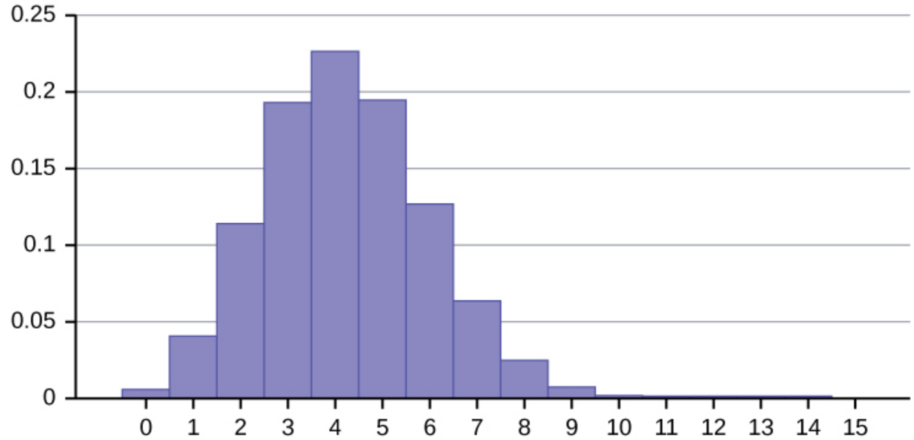
Problem:

The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.

a. Sketch a graph of the probability distribution of X.
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Find the probability that more than five people in the sample are literate. Is it is more likely that three people or four people are literate.

Solution:

a. $X \sim B(15, 0.281)$



b. i. Mean = μ = np = 15(0.281) = 4.215

ii. Standard Deviation = σ = $\sqrt{npq}$ = $\sqrt{15(0.281)(0.719)}$ = 1.7409

c. $P(x > 5) = 1 – P(x \le 5)$ = 1 – binomcdf(15, 0.281, 5) = 1 – 0.7754 = 0.2246

$P(x = 3)$ = binompdf(15, 0.281, 3) = 0.1927

$P(x = 4)$ = binompdf(15, 0.281, 4) = 0.2259

It is more likely that four people are literate that three people are.

Glossary

Binomial Experiment

a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, *n*.
2. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter *p* denotes the probability of a success on one trial, and *q* denotes the probability of a failure on one trial.
3. The *n* trials are independent and are repeated using identical conditions.

Bernoulli Trials

an experiment with the following characteristics:

1. There are only two possible outcomes called “success” and “failure” for each trial.
2. The probability *p* of a success is the same for any trial (so the probability $q = 1 − p$ of a failure is the same for any trial).

Binomial Probability Distribution

a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, *n*, of independent trials. “Independent” means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same

conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Introduction class="introduction"

The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

Note:
Note
The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated as **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

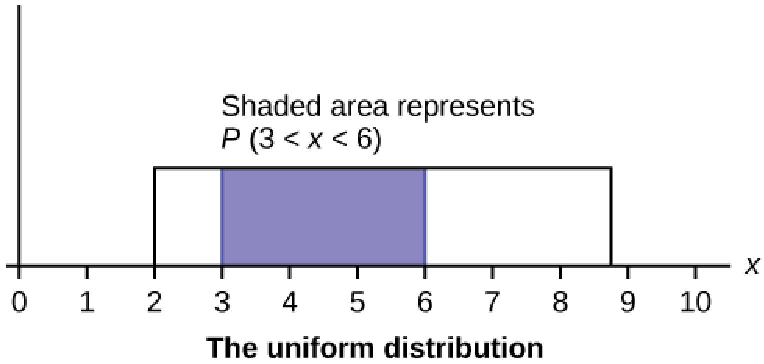
Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

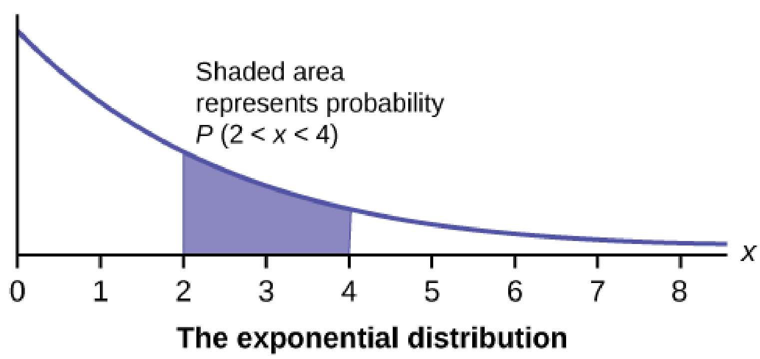
We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter and the next, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

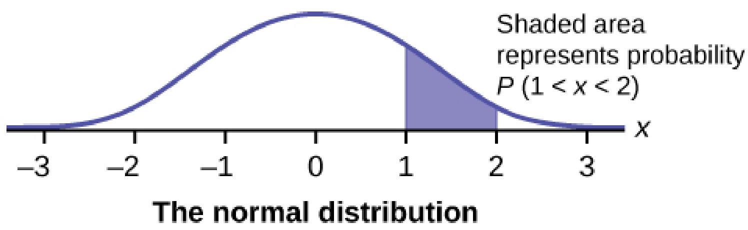


The graph shows a Uniform Distribution with the area between x = 3 and x = 6 shaded to represent the probability that the value of the random variable X is in the interval between three and six.



The graph shows an Exponential Distribution with the area between x = 2 and x = 4 shaded to represent the probability that the value of the random

variable X is in the interval between two and four.



The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

Glossary

Uniform Distribution
: a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Notation: $X \sim U(a,b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \le x \le b$. The cumulative distribution is $P(X \le x) = \frac{x-a}{b-a}$.

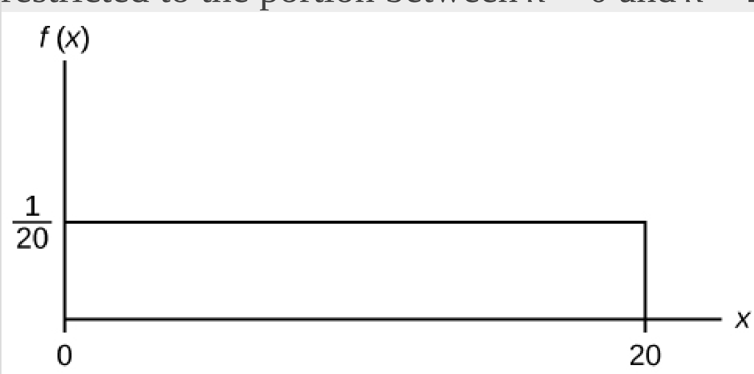
Exponential Distribution
: a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital;

the notation is $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Continuous Probability Functions

We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. **For continuous probability distributions, PROBABILITY = AREA.**

Example:
Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. x = a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.

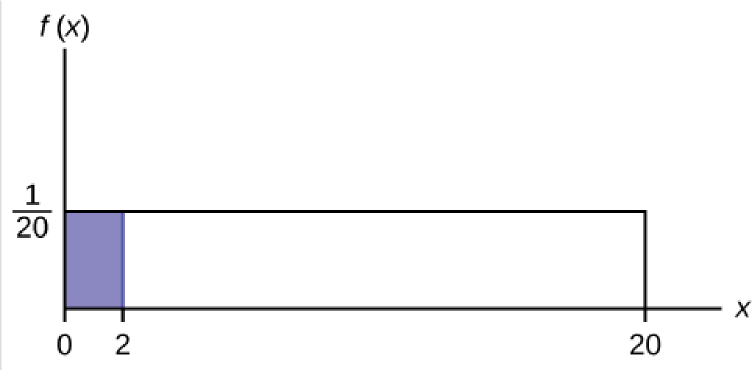


$f(x) = \frac{1}{20}$ **for** $0 \leq x \leq 20$.
The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.
The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

Equation:

$$\text{AREA} = 20\left(\frac{1}{20}\right) = 1$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $0 < x < 2$.



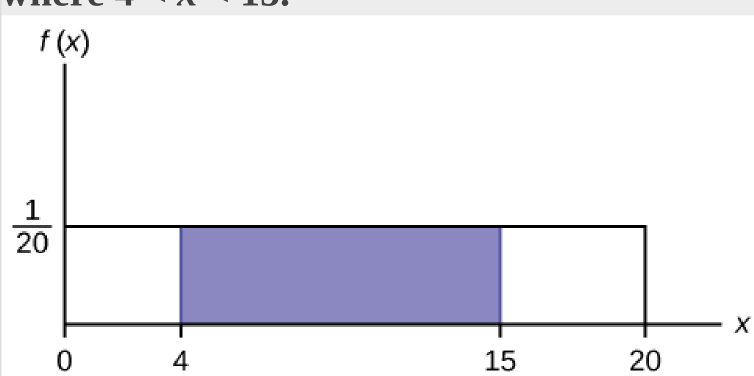
$$\text{AREA} = (2-0)\left(\frac{1}{20}\right) = 0.1$$

$$(2-0) = 2 = \text{base of a rectangle}$$

Note:
Reminder
area of a rectangle = (base)(height).

The area corresponds to a probability. The probability that *x* is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

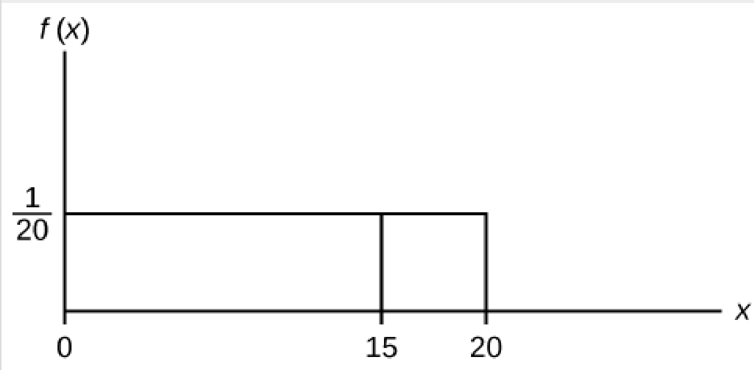
Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the *x*-axis where $4 < x < 15$.



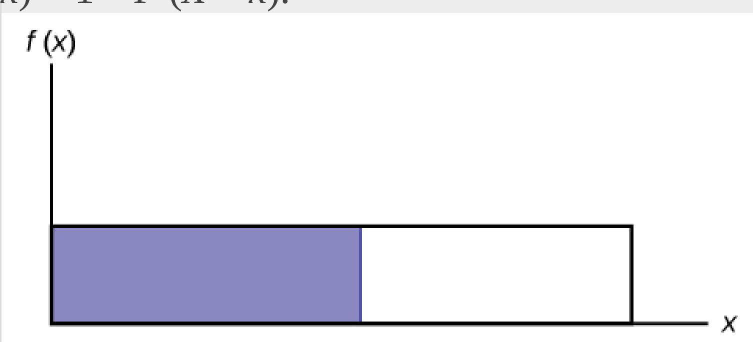
$$\text{AREA} = (15-4)\left(\frac{1}{20}\right) = 0.55$$
$$(15-4) = 11 = \text{the base of a rectangle}$$

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find $P(x = 15)$. On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or zero width). Therefore, $P(x = 15) = (\text{base})(\text{height}) = (0)\left(\frac{1}{20}\right) = 0$

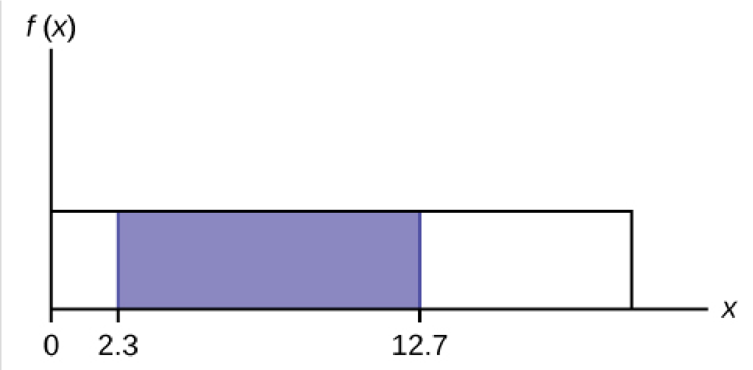


$P(X \leq x)$, which can also be written as $P(X < x)$ for continuous distributions, is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can also use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 – P(X < x)$.



Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \leq x \leq 20$.

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.



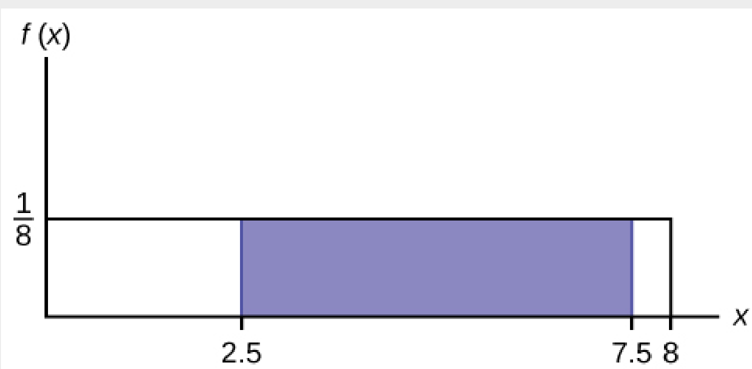
$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$

Note:
Try It
Exercise:

Problem:

Consider the function $f(x) = \frac{1}{8}$ for $0 \le x \le 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

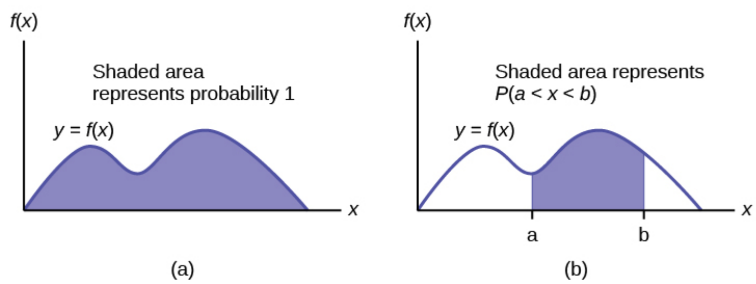
Solution:



$P(2.5 < x < 7.5) = 0.625$

Chapter Review

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(a < x < b)$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.



The cumulative distribution function (cdf) of X is defined by $P(X \leq x)$. It is a function of x that gives the probability that the random variable is less than or equal to x.

Formula Review

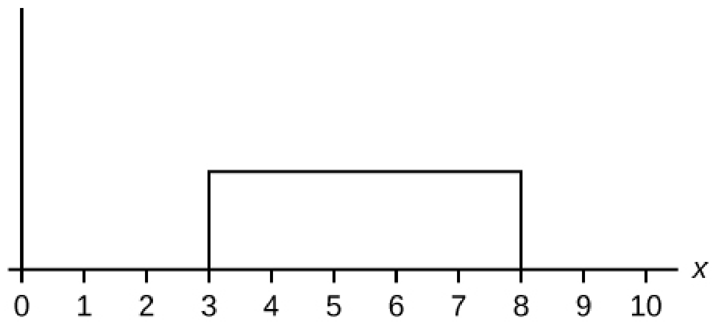
Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

Cumulative distribution function (cdf): $P(X \leq x)$

Exercise:

Problem: Which type of distribution does the graph illustrate?

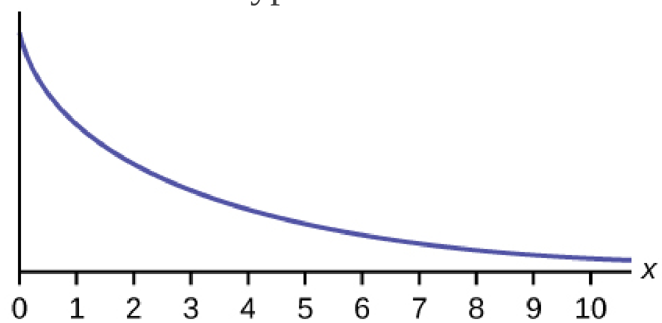


Solution:

Uniform Distribution

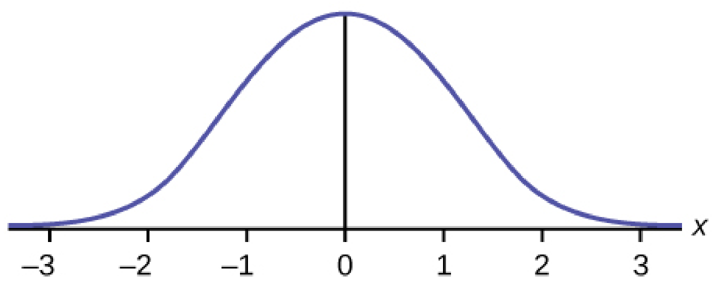
Exercise:

Problem: Which type of distribution does the graph illustrate?



Exercise:

Problem: Which type of distribution does the graph illustrate?

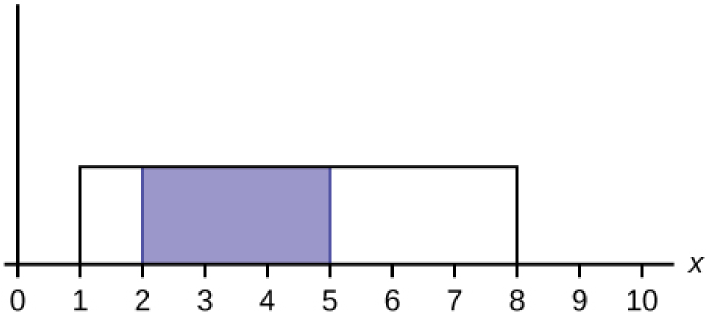


Solution:

Normal Distribution

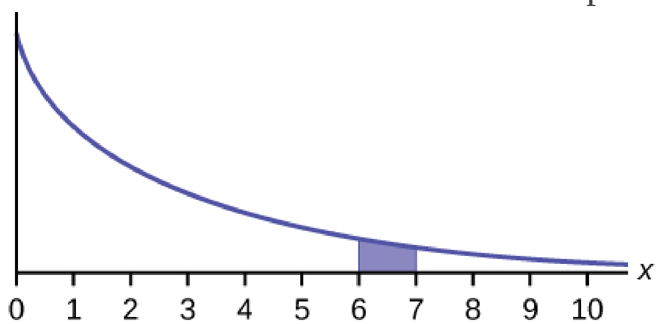
Exercise:

Problem: What does the shaded area represent? $P($___$< x <$ ___$)$



Exercise:

Problem: What does the shaded area represent? $P($___$< x <$ ___$)$



Solution:

$P(6 < x < 7)$

Exercise:

Problem:

For a continuous probablity distribution, $0 \leq x \leq 15$. What is $P(x > 15)$?

Exercise:

Problem:

What is the area under $f(x)$ if the function is a continuous probability density function?

Solution:

one

Exercise:

Problem:

For a continuous probability distribution, $0 \le x \le 10$. What is $P(x = 7)$?

Exercise:

Problem:

A **continuous** probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?

Solution:

zero

Exercise:

Problem:

$f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \le x \le 5$. What is $P(x < 0)$?

Exercise:

Problem:

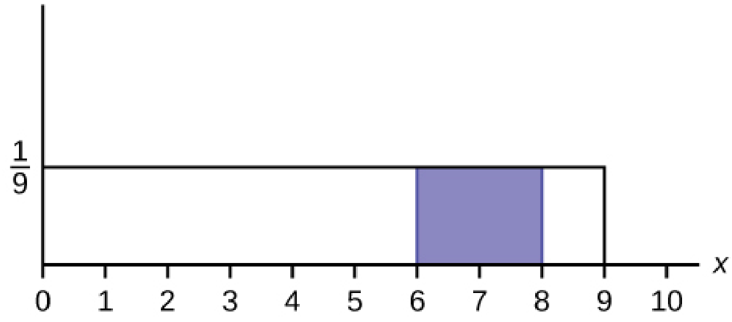
$f(x)$, a continuous probability function, is equal to $\frac{1}{12}$, and the function is restricted to $0 \le x \le 12$. What is $P\ (0 < x < 12)$?

Solution:

one

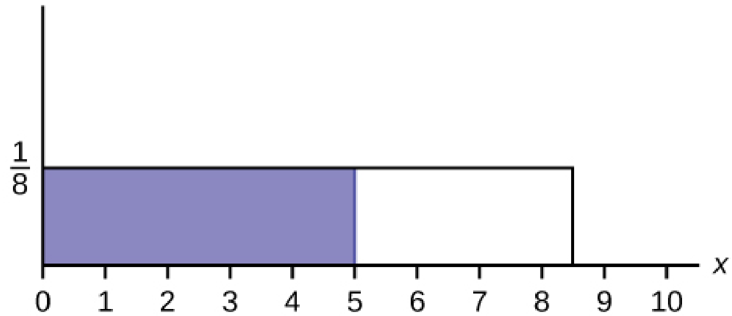
Exercise:

Problem: Find the probability that x falls in the shaded area.



Exercise:

Problem: Find the probability that *x* falls in the shaded area.

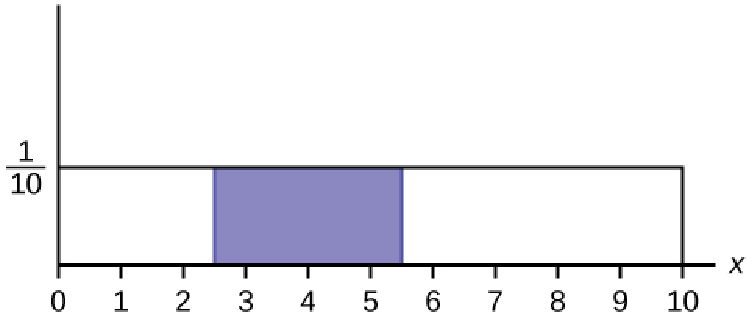


Solution:

0.625

Exercise:

Problem: Find the probability that *x* falls in the shaded area.



Exercise:

Problem:

$f(x)$, a continuous probability function, is equal to $\frac{1}{3}$ and the function is restricted to $1 \leq x \leq 4$. Describe $P\left(x > \frac{3}{2}\right)$.

Solution:

The probability is equal to the area from $x = \frac{3}{2}$ to $x = 4$ above the x-axis and up to $f(x) = \frac{1}{3}$.

Homework

For each probability and percentile problem, draw the picture.

Exercise:

Problem:

Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer "yes" or "no." You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.

a. What part of the experiment will yield discrete data?
b. What part of the experiment will yield continuous data?

Exercise:

Problem:

When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

Solution:

Age is a measurement, regardless of the accuracy used.

Introduction
class="introduction"

If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

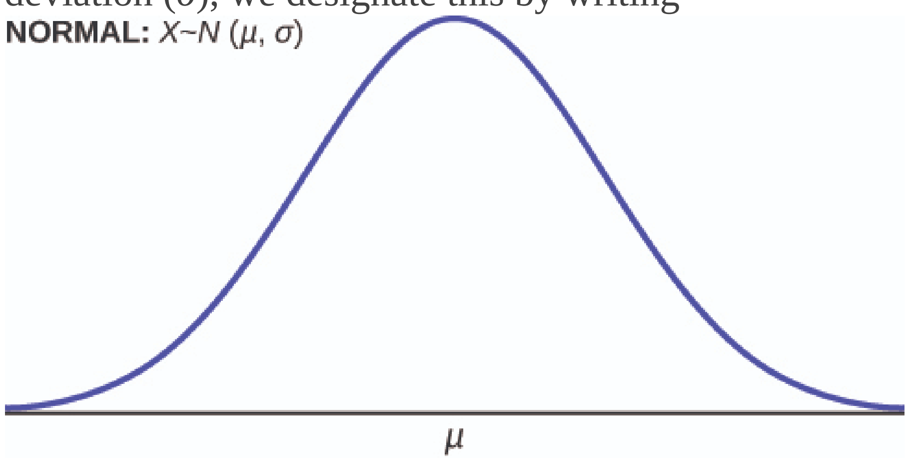
- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these

include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures): the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing



The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions.

The curve is symmetric about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric

about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

Note:
Collaborative Classroom Activity
Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

$X \sim N(\mu, \sigma)$

μ = the mean σ = the standard deviation

Glossary

Normal Distribution
    a continuous random variable (RV) with pdf $f(x)$ =

$$\frac{1}{\sigma\sqrt{2\pi}}\ e^{\frac{-(x-\mu)^{2}}{2\sigma^{2}}}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called *z*-scores**. **A *z*-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$

The *z*-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x in the given equation comes from a normal distribution with mean μ and standard deviation σ.

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score is:

Equation:

$$z = \frac{x - \mu}{\sigma}$$

The *z*-score tells you how many standard deviations the value *x* is above (to the right of) or below (to the left of) the mean, *μ*. Values of x that are larger than the mean have positive *z*-scores, and values of x that are smaller than the mean have negative *z*-scores. If x equals the mean, then x has a *z*-score of zero.

Example:

Suppose $X \sim N(5, 6)$. This says that X is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$.

Then:
Equation:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two standard deviations** (2σ) above or to the right of the mean $\mu = 5$.
Notice that: 5 + (2)(6) = 17 (The pattern is $\mu + z\sigma = x$)
Now suppose $x = 1$. Then: $z = \frac{x-\mu}{\sigma} = \frac{1-5}{6} = -0.67$ (rounded to two decimal places)
This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: 5 + (–0.67)(6) is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)
Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.

Note:
Try It
Exercise:

Problem: What is the z-score of x, when $x = 1$ and $X \sim N(12,3)$?

Solution:

$z = \frac{1-12}{3} \approx -3.67$

Example:
Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of

two pounds. $X \sim N(5, 2)$. Fill in the blanks.

Exercise:

Problem:

a. Suppose a person **lost** ten pounds in a month. The z-score when x = 10 pounds is z = 2.5 (verify). This z-score tells you that x = 10 is _________ standard deviations to the _________ (right or left) of the mean ______ (What is the mean?).

Solution:

a. This z-score tells you that x = 10 is **2.5** standard deviations to the **right** of the mean **five**.

Exercise:

Problem:

b. Suppose a person **gained** three pounds (a negative weight loss). Then z = ___________. This z-score tells you that x = –3 is _________ standard deviations to the ___________ (right or left) of the mean.

Solution:

b. z = **–4**. This z-score tells you that x = –3 is **four** standard deviations to the **left** of the mean.

Exercise:

Problem:

c. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If x = 17, then z = 2. (This was previously shown.) If y = 4, what is z?

Solution:

c. $z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2$ where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.

The *z*-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Note:
Try It
Exercise:

Problem: Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is –1.5. This score tells you that $x = 10$ is ______ standard deviations to the ______(right or left) of the mean______(What is the mean?).

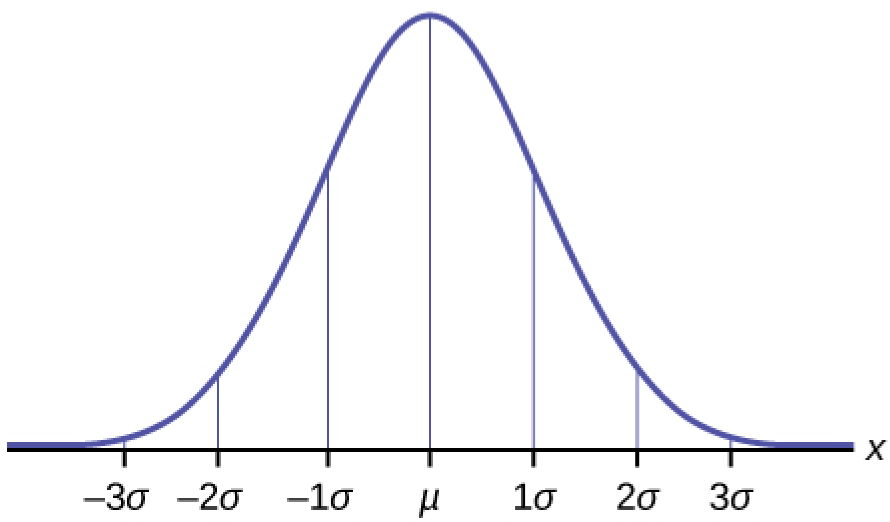
Solution:

1.5, left, 16

The Empirical Rule
If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** states the following:

- About 68% of the x values lie between $–1\sigma$ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between $–2\sigma$ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between $–3\sigma$ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and $–1\sigma$ are +1 and –1, respectively.
- The z-scores for $+2\sigma$ and $–2\sigma$ are +2 and –2, respectively.
- The z-scores for $+3\sigma$ and $–3\sigma$ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.



Example:
The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Exercise:

Problem:

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when x = 168 cm is z = ________. This z-score tells you that x = 168 is _________ standard deviations to the _________ (right or left) of the mean _____ (What is the mean?).

Solution:

a. –0.32, 0.32, left, 170

Exercise:

Problem:

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of z = 1.27. What is the male's height? The z-score (z = 1.27) tells you that the male's height is _________ standard deviations to the ___________ (right or left) of the mean.

Solution:

b. 177.98 cm, 1.27, right

Note:
Try It
Exercise:

Problem:

Use the information in [link] to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when x = 176 cm is z = ________. This z-score tells you that x = 176 cm is _________ standard

deviations to the _________ (right or left) of the mean ______ (What is the mean?).

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of z = –2. What is the male's height? The z-score (z = –2) tells you that the male's height is _________ standard deviations to the ___________ (right or left) of the mean.

Solution:
Try It Solutions

Solve the equation $z = \frac{x-\mu}{\sigma}$ for x. $x = \mu + (z)(\sigma)$

a. $z = \frac{176-170}{6.28} \approx 0.96$, This z-score tells you that x = 176 cm is 0.96 standard deviations to the right of the mean 170 cm.
b. X = 157.44 cm, The z-score(z = –2) tells you that the male's height is two standard deviations to the left of the mean.

Example:
Exercise:

Problem:

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the z-scores for x = 160.58 cm and y = 162.85 cm. Interpret each z-score. What can you say about x = 160.58 cm and y = 162.85 cm as they compare to their respective means and standard deviations?

Solution:

The z-score for x = -160.58 is z = –1.5.
The z-score for y = 162.85 is z = –1.5.
Both x = 160.58 and y = 162.85 deviate the same number of standard deviations from their respective means and in the same direction.

Note:
Try It
Exercise:

Problem:

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean μ = 496 and a standard deviation σ = 114. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for x_1 = 325 and x_2 = 366.21. Interpret each z-score. What can you say about x_1 = 325 and x_2 = 366.21 as they compare to their respective means and standard deviations?

Solution:

The z-score for x_1 = 325 is z_1 = –1.5.

The z-score for x_2 = 366.21 is z_2 = –1.14.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example:

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie within one standard deviation of the mean. Therefore, about 68% of the x values lie between $–1\sigma = (–1)(6) = –6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values 50 – 6 = 44 and 50 + 6 = 56 are within one standard deviation from the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between $–2\sigma = (–2)(6) = –12$ and $2\sigma = (2)(6) = 12$. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations from the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 95% of the x values lie between $–3\sigma = (–3)(6) = –18$ and $3\sigma = (3)(6) = 18$ from the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Note:

Try It

Exercise:

Problem:

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Solution:

between 20 and 30.

Example:
Exercise:

Problem:

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let *Y* = the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

a. About 68% of the *y* values lie between what two values? These values are ________________. The *z*-scores are ________________, respectively.
b. About 95% of the *y* values lie between what two values? These values are ________________. The *z*-scores are ________________ respectively.
c. About 99.7% of the *y* values lie between what two values? These values are ________________. The *z*-scores are ________________, respectively.

Solution:

a. About 68% of the values lie between 166.02 cm and 178.7 cm. The *z*-scores are –1 and 1.
b. About 95% of the values lie between 159.68 cm and 185.04 cm. The *z*-scores are –2 and 2.
c. About 99.7% of the values lie between 153.34 cm and 191.38 cm. The *z*-scores are –3 and 3.

Note:
Try It
Exercise:

Problem:

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the *y* values lie between what two values? These values are __________________. The *z*-scores are __________________, respectively.
b. About 95% of the *y* values lie between what two values? These values are __________________. The *z*-scores are __________________, respectively.
c. About 99.7% of the *y* values lie between what two values? These values are __________________. The *z*-scores are __________________, respectively.

Solution:

a. About 68% of the values lie between the values 41 and 63. The *z*-scores are –1 and 1, respectively.
b. About 95% of the values lie between the values 30 and 74. The *z*-scores are –2 and 2, respectively.
c. About 99.7% of the values lie between the values 19 and 85. The *z*-scores are –3 and 3, respectively.

References

"Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

"The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores."

London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

“2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).

“Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Chapter Review

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Formula Review

z = a standardized value (z-score)

mean = 0; standard deviation = 1

To find the k^{th} percentile of X when the z-scores is known:
$k = \mu + (z)\sigma$

z-score: $z = \frac{x - \mu}{\sigma}$

Z = the random variable for z-scores

Exercise:

Problem:

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. X = _____________.

Solution:

ounces of water in a bottle

Exercise:

Problem:

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Exercise:

Problem: $X \sim N(1, 2)$

σ = ________

Solution:

2

Exercise:

Problem:

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. X = _______________.

Exercise:

Problem: $X \sim N(–4, 1)$

What is the median?

Solution:

–4

Exercise:

Problem: $X \sim N(3, 5)$

σ = ________

Exercise:

Problem: $X \sim N(–2, 1)$

μ = ________

Solution:

–2

Exercise:

Problem: What does a z-score measure?

Exercise:

Problem:

What does standardizing a normal distribution do to the mean?

Solution:

The mean becomes zero.

Exercise:

Problem:

Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

Exercise:

Problem:

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Solution:

$z = 2$

Exercise:

Problem:

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

Exercise:

Problem:

What is the z-score of $x = –2$, if it is 2.78 standard deviations to the right of the mean?

Solution:

$z = 2.78$

Exercise:

Problem:

What is the *z*-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

Exercise:

Problem: Suppose $X \sim N(2, 6)$. What value of *x* has a *z*-score of three?

Solution:

$x = 20$

Exercise:

Problem:

Suppose $X \sim N(8, 1)$. What value of *x* has a *z*-score of –2.25?

Exercise:

Problem: Suppose $X \sim N(9, 5)$. What value of *x* has a *z*-score of –0.5?

Solution:

$x = 6.5$

Exercise:

Problem:

Suppose $X \sim N(2, 3)$. What value of *x* has a *z*-score of –0.67?

Exercise:

Problem:

Suppose $X \sim N(4, 2)$. What value of *x* is 1.5 standard deviations to the left of the mean?

Solution:

$x = 1$

Exercise:

Problem:

Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

Exercise:

Problem:

Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

Solution:

$x = 1.97$

Exercise:

Problem: Suppose $X \sim N(–1, 2)$. What is the z-score of $x = 2$?

Exercise:

Problem: Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

Solution:

$z = –1.67$

Exercise:

Problem: Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

Exercise:

Problem:

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

Solution:

$z \approx -0.33$

Exercise:

Problem:

In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is _____ standard deviations to the _____ (right or left) of the mean.

Exercise:

Problem:

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

0.67, right

Exercise:

Problem:

In a normal distribution, $x = -2$ and $z = 6$. This tells you that $x = -2$ is _____ standard deviations to the _____ (right or left) of the mean.

Exercise:

Problem:

In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

3.14, left

Exercise:

Problem:

In a normal distribution, *x* = 6 and *z* = –1.7. This tells you that *x* = 6 is ____ standard deviations to the ____ (right or left) of the mean.

Exercise:

Problem:

About what percent of *x* values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Solution:

about 68%

Exercise:

Problem:

About what percent of the *x* values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Exercise:

Problem:

About what percent of *x* values lie between the second and third standard deviations (both sides)?

Solution:

about 4%

Exercise:

Problem:

Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Exercise:

Problem:

Suppose $X \sim N(–3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution(i.e., –3).

Solution:

between –5 and –1

Exercise:

Problem:

Suppose $X \sim N(–3, 1)$. Between what x values does 34.14% of the data lie?

Exercise:

Problem:

About what percent of x values lie between the mean and three standard deviations?

Solution:

about 50%

Exercise:

Problem:

About what percent of x values lie between the mean and one standard deviation?

Exercise:

Problem:

About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

Solution:

about 27%

Exercise:

Problem:

About what percent of x values lie betwween the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

Exercise:

Problem:

Define the random variable X in words. X = ________________.

Solution:

The lifetime of a Sunshine CD player measured in years.

Exercise:

Problem: X ~ _____(_____,_____)

Homework

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Exercise:

Problem: What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

Exercise:

Problem:

What is the *z*-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

Solution:

c

Exercise:

Problem:

The length of time to find it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution..

III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only
d. I, II, and III

Exercise:

Problem:

The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.

Solution:

a. Use the z-score formula. $z = –0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.
b. Use the z-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
c. Height = 79 + 3.5(3.89) = 92.615 inches, which is taller than 7 feet, 8 inches. There are very few NBA players this tall so the answer is no, not likely.

Exercise:

Problem:

The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.

a. Calculate the *z*-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

Exercise:

Problem:

Kyle's doctor told him that the *z*-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If X = a systolic blood pressure score then $X \sim N(125, 14)$.

a. Which answer(s) **is/are** correct?

   i. Kyle's systolic blood pressure is 175.
   ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
   iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.

b. Calculate Kyle's blood pressure.

Solution:

a. iv
b. Kyle's blood pressure is equal to 125 + (1.75)(14) = 149.5.

Exercise:

Problem:

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean μ = 10.2 kg and standard deviation σ = 0.8 kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the z-scores that correspond to the following weights and interpret them.

a. 11 kg
b. 7.9 kg
c. 12.2 kg

Exercise:

Problem:

In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean μ = 520 and standard deviation σ = 115.

a. Calculate the z-score for an SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored

700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

Solution:

Let X = an SAT math score and Y = an ACT math score.

a. $X = 720$ $\frac{720-520}{15} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.

b. $z = 1.5$
The math SAT score is $520 + 1.5(115) \approx 692.5$. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.

c. $\frac{X-\mu}{\sigma} = \frac{700-514}{117} \approx 1.59$, the z-score for the SAT. $\frac{Y-\mu}{\sigma} = \frac{30-21}{5.3} \approx 1.70$, the z-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher z-score).

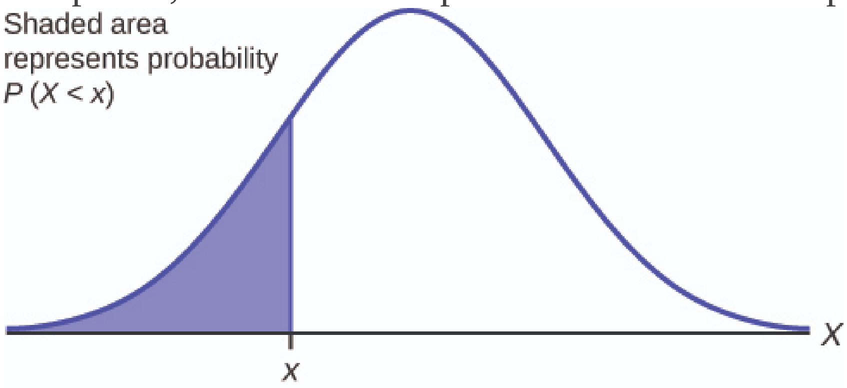
Glossary

Standard Normal Distribution
: a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

z-score
: the linear transformation of the form $z = \frac{x-\mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.



The area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x)$ = **Area to the left** of the vertical line through x. $P(X < x) = 1 – P(X < x)$ = **Area to the right** of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.

Note:
NOTE
To calculate the probability, use the probability tables provided in [link] without the use of technology. The tables include instructions for how to use them.

Example:
If the area to the left is 0.0228, then the area to the right is 1 – 0.0228 = 0.9772.

Note:
Try It
Exercise:

Problem:

If the area to the left of x is 0.012, then what is the area to the right?

Solution:

$1 - 0.012 = 0.988$

Example:
The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

Exercise:

Problem:

a. Find the probability that a randomly selected student scored more than 65 on the exam.

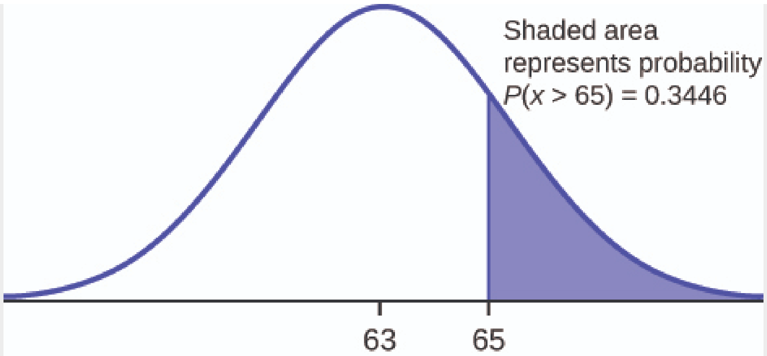
Solution:

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$.

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$



The probability that any student selected at random scores more than 65 is 0.3446.

Note:
Go into `2nd DISTR`.
After pressing `2nd DISTR`, press `2:normalcdf`.
The syntax for the instructions are as follows:
normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1`, the `EE` key (a 2nd key) and then `99`. Or, you can enter `10^99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Note:
Historical Note
The TI probability program calculates a *z*-score and then the probability from the *z*-score. Before technology, the *z*-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the *z*-score was used.

You calculate the *z*-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65 - 63}{5} = 0.4$$

Area to the left is 0.6554.

$P(x > 65) = P(z > 0.4) = 1 – 0.6554 = 0.3446$

Note:
Find the percentile for a student scoring 65:
*Press `2nd Distr`
*Press `2:normalcdf`(
*Enter lower bound, upper bound, mean, standard deviation followed by)
*Press `ENTER`.
For this Example, the steps are
`2nd Distr`
`2:normalcdf`(65,1,2nd EE,99,63,5) `ENTER`
The probability that a selected student scored more than 65 is 0.3446.

Exercise:

Problem:

b. Find the probability that a randomly selected student scored less than 85.

Solution:

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

normalcdf(0,85,63,5) = 1 (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Exercise:

Problem:

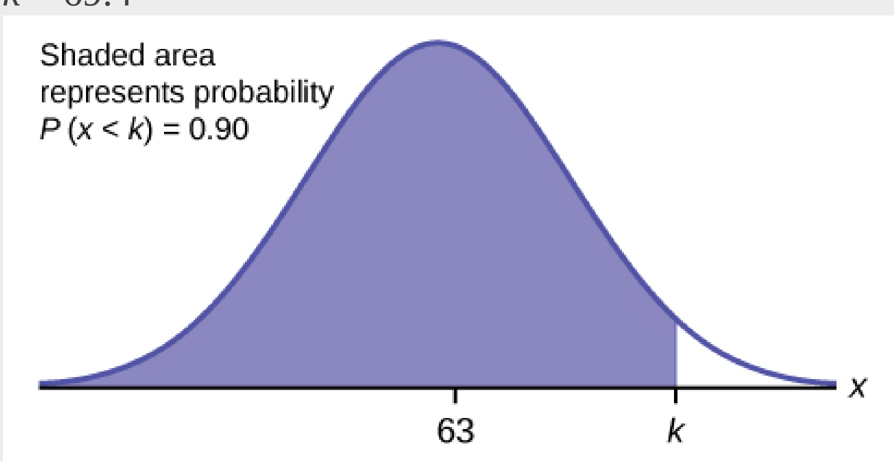
c. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).

Solution:

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a **critical value**.

$k = 69.4$



The 90^{th} percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

Note:
`invNorm` in `2nd DISTR`. invNorm(area to the left, mean, standard deviation)
For this problem, invNorm(0.90,63,5) = 69.4

Exercise:

Problem:

d. Find the 70^{th} percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution:

d. Find the 70^{th} percentile.

Draw a new graph and label it appropriately. k = 65.6

The 70^{th} percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

invNorm(0.70,63,5) = 65.6

Note:
Try It
Exercise:

Problem:

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

Solution:

normalcdf(0,65,68,3) = 0.1587

Example:

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Exercise:

Problem:

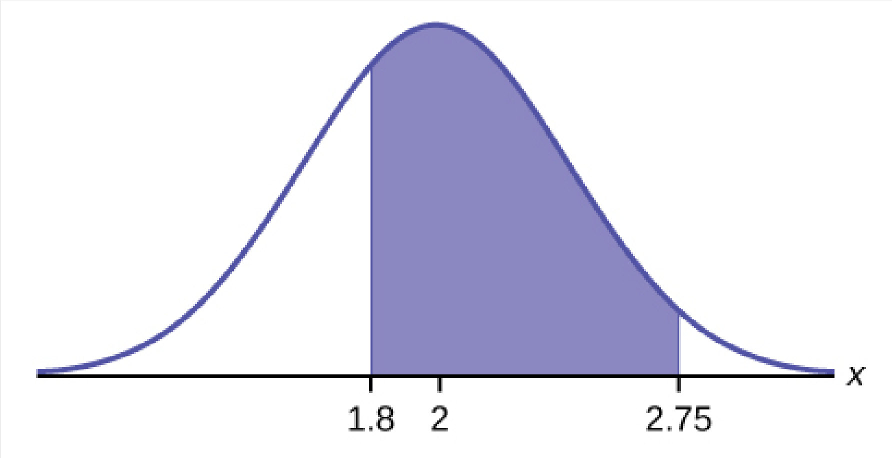
a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution:

a. Let X = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$



normalcdf(1.8,2.75,2,0.5) = 0.5886

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

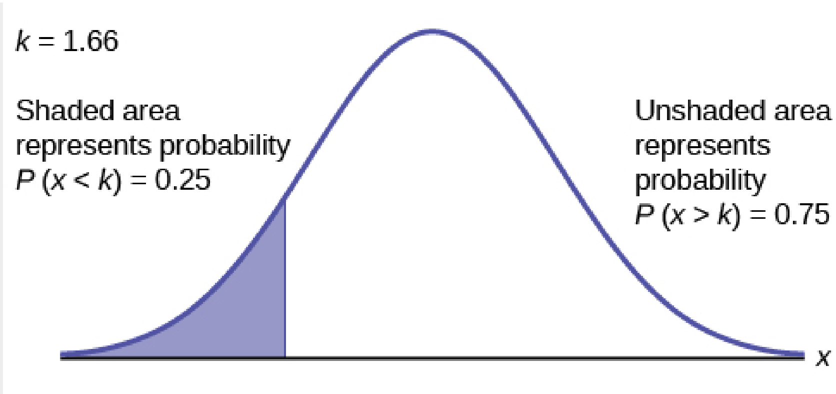
Exercise:

Problem:

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution:

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25^{th} percentile,** k, where $P(x < k) = 0.25$.



invNorm(0.25,2,0.5) = 1.66

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Note:
Try It
Exercise:

Problem:

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Solution:

normalcdf(66,70,68,3) = 0.4950

Example:
In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

Exercise:

Problem:

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Solution:

a. normalcdf(23,64.7,36.9,13.9) = 0.8186

Exercise:

Problem:

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Solution:

b. normalcdf($–10^{99}$,50.8,36.9,13.9) = 0.8413

Exercise:

Problem:

c. Find the 80^{th} percentile of this distribution, and interpret it in a complete sentence.

Solution:

c.

- invNorm(0.80,36.9,13.9) = 48.6
- The 80^{th} percentile is 48.6 years.
- 80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Note:
Try It
Use the information in [link] to answer the following questions.
Exercise:

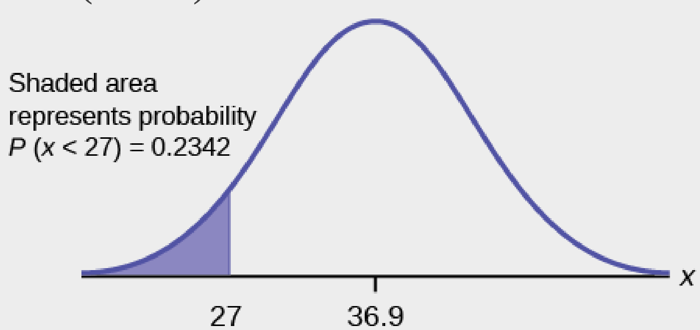
Problem:

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old.

Solution:

Let X = a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

a. To find the 30th percentile, find k such that $P(x < k) = 0.30$.
invNorm(0.30, 36.9, 13.9) = 29.6 years
Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years.
b. Find $P(x < 27)$



normalcdf(0,27,36.9,13.9) = 0.2342
(Note that normalcdf(-10^{99},27,36.9,13.9) = 0.2382. The two answers differ only by 0.0040.)

Example:

In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information, answer the following questions (round answers to one decimal place).

Exercise:

Problem: a. Calculate the interquartile range (*IQR*).

Solution:

a.

- $IQR = Q_3 – Q_1$
- Calculate $Q_3 = 75^{th}$ percentile and $Q_1 = 25^{th}$ percentile.
- invNorm(0.75,36.9,13.9) = Q_3 = 46.2754
- invNorm(0.25,36.9,13.9) = Q_1 = 27.5246
- $IQR = Q_3 – Q_1 = 18.8$

Exercise:

Problem:

b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution:

b.

- Find k where $P(x \geq k) = 0.40$ ("At least" translates to "greater than or equal to.")
- 0.40 = the area to the right.
- Area to the left = 1 – 0.40 = 0.60.
- The area to the left of k = 0.60.
- invNorm(0.60,36.9,13.9) = 40.4215.

- k = 40.4.
- Forty percent of the ages that range from 13 to 55+ are at least 40.4 years.

Note:
Try It
Exercise:

Problem:

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean μ = 81 points and standard deviation σ = 15 points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Solution:

a. Q_1 = 25th percentile = invNorm(0.25,81,15) = 70.9
   Q_3 = 75th percentile = invNorm(0.75,81,15) = 91.1
b. The middle 50% of the scores are between 70.9 and 91.1.

Example:
A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

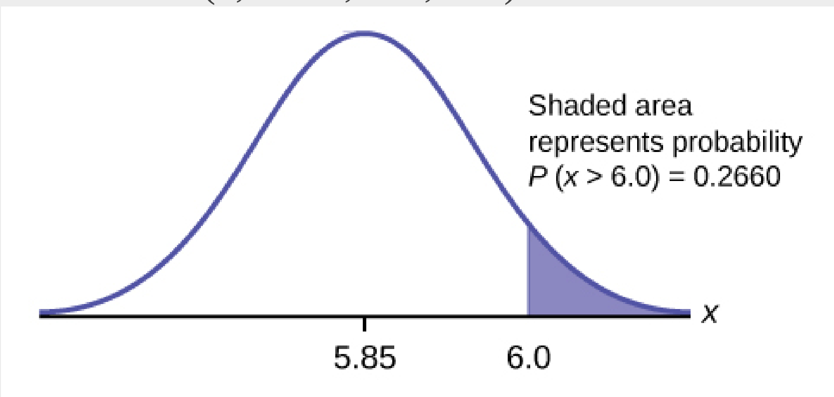
Exercise:

Problem:

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.

Solution:

a. normalcdf(6,10^99,5.85,0.24) = 0.2660



Exercise:

Problem:

b. The middle 20% of mandarin oranges from this farm have diameters between _______ and _______.

Solution:

b.

- 1 – 0.20 = 0.80
- The tails of the graph of the normal distribution each have an area of 0.40.
- Find *k1*, the 40th percentile, and *k2*, the 60th percentile (0.40 + 0.20 = 0.60).
- *k1* = invNorm(0.40,5.85,0.24) = 5.79 cm
- *k2* = invNorm(0.60,5.85,0.24) = 5.91 cm

Exercise:

Problem:

c. Find the 90^{th} percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Solution:

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.16 cm.

Note:
Try It
Exercise:

Problem: Using the information from [link], answer the following:

a. The middle 40% of mandarin oranges from this farm are between _______ and _______.
b. Find the 16^{th} percentile and interpret it in a complete sentence.

Solution:

a. The middle area = 0.40, so each tail has an area of 0.30.

   1 – 0.40 = 0.60

   The tails of the graph of the normal distribution each have an area of 0.30.

   Find *k1*, the 30^{th} percentile and *k2*, the 70^{th} percentile (0.40 + 0.30 = 0.70).

   k1 = invNorm(0.30,5.85,0.24) = 5.72 cm

$k2$ = invNorm(0.70,5.85,0.24) = 5.98 cm

b. invNorm(0.16, 5.85, 0.24) = 5.61; 16% of mandarin oranges from this farm have diameter 5.61 cm or less.

References

"Naegele's rule." Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).

"403: NUMMI." Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radio-archives/episode/403/nummi (accessed May 14, 2013).

"Scratch-Off Lottery Ticket Playing Tips." WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/public/department40.cfm (accessed May 14, 2013).

"Smart Phone Users, By The Numbers." Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

"Facebook Statistics." Statistics Brain. Available online at http://www.statisticbrain.com/facebook-statistics/(accessed May 14, 2013).

Chapter Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.

Standard Normal Distribution: $Z \sim N(0, 1)$.

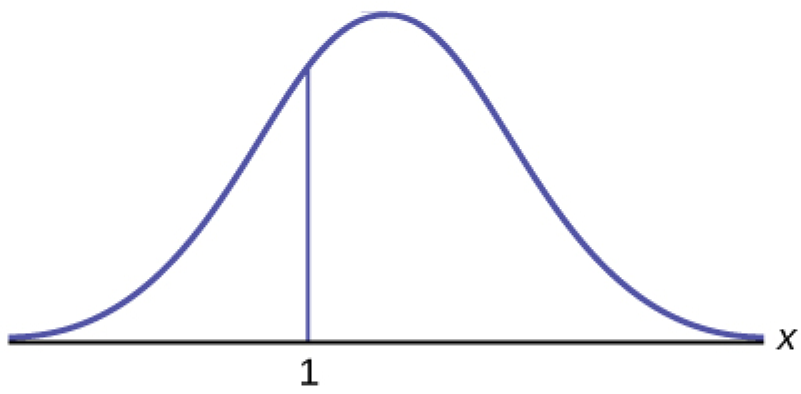
Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)

Calculator function for the k^{th} percentile: k = invNorm (area to the left of k, mean, standard deviation)

Exercise:

Problem:

How would you represent the area to the left of one in a probability statement?

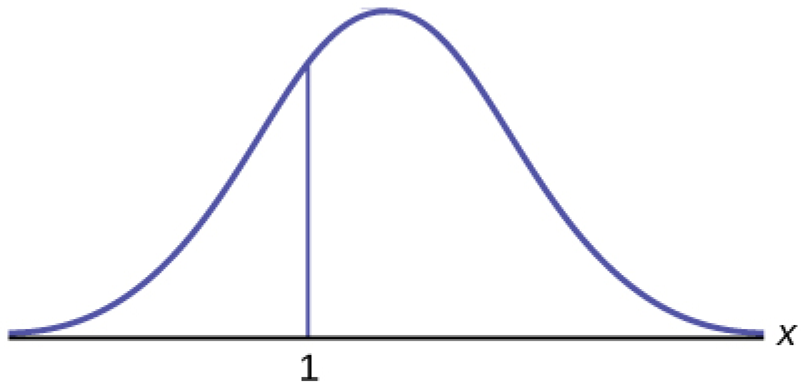


Solution:

$P(x < 1)$

Exercise:

Problem: What is the area to the right of one?



Exercise:

Problem: Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

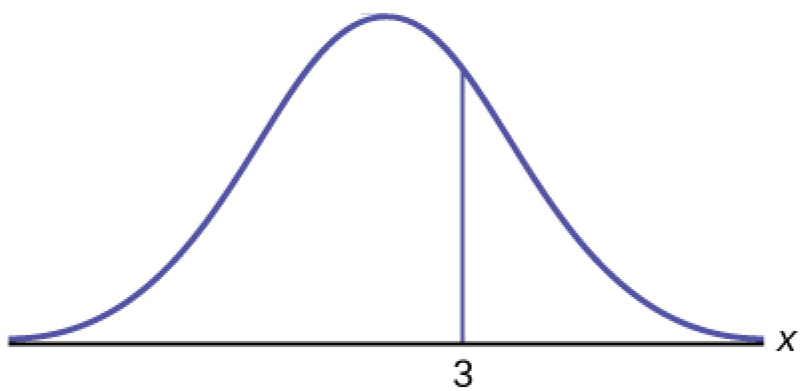
Solution:

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

Exercise:

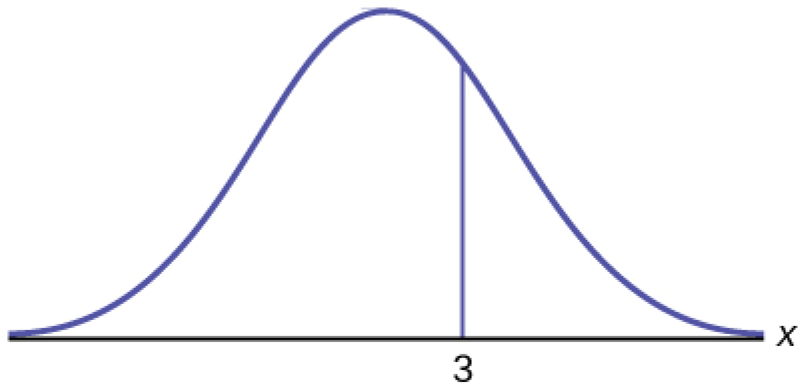
Problem:

How would you represent the area to the left of three in a probability statement?



Exercise:

Problem: What is the area to the right of three?



Solution:

$1 - P(x < 3)$ or $P(x > 3)$

Exercise:

Problem:

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Exercise:

Problem:

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Solution:

$1 - 0.543 = 0.457$

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

Exercise:

Problem: Find the probability that $x > 56$.

Exercise:

Problem: Find the probability that $x < 30$.

Solution:

0.0013

Exercise:

Problem: Find the 80^{th} percentile.

Exercise:

Problem: Find the 60^{th} percentile.

Solution:

56.03

Exercise:

Problem: $X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise:

Problem: $X \sim N(–3, 4)$

Find the probability that x is between one and four.

Solution:

0.1186

Exercise:

Problem: $X \sim N(4, 5)$

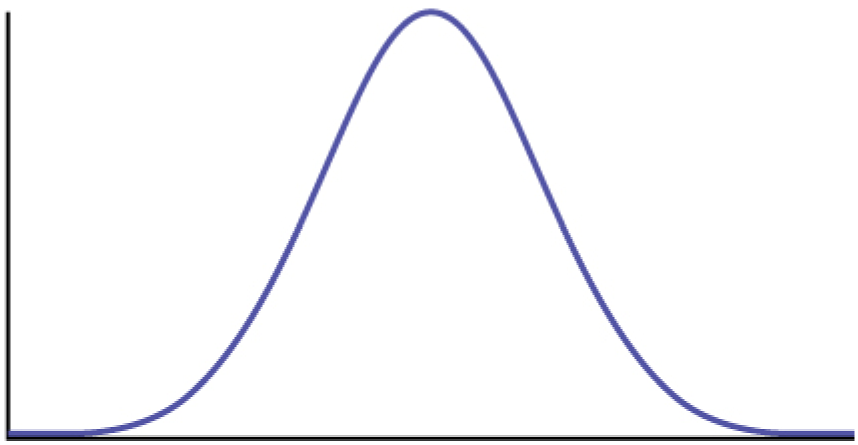
Find the maximum of *x* in the bottom quartile.

Exercise:

Problem:

Use the following information to answer the next three exercise: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



b. $P(0 < x <$ _____________$) =$ ____________ (Use zero for the minimum value of *x*.)

Solution:

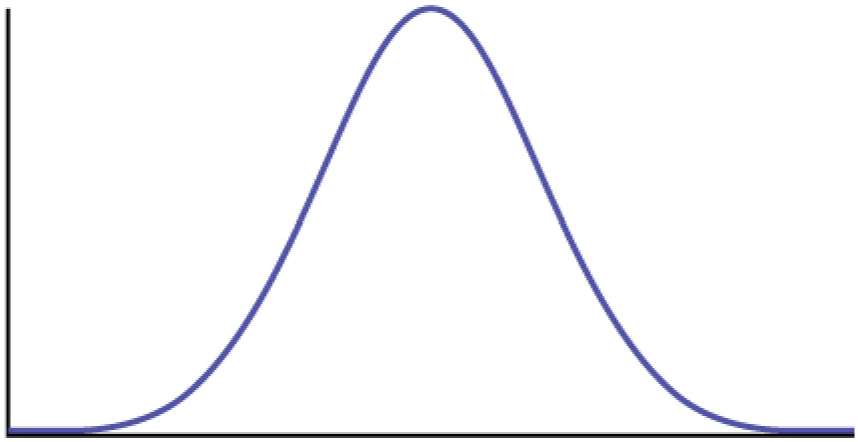
a. Check student's solution.
b. 3, 0.1979

Exercise:

Problem:

Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



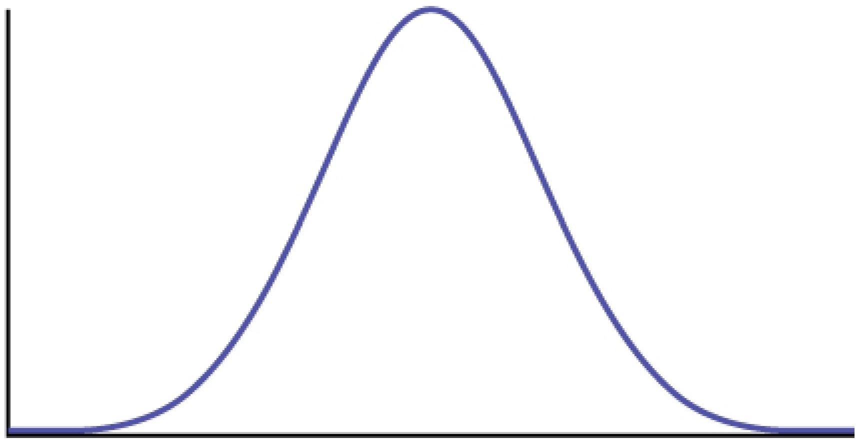
b. $P(\_\_\_\_\_\_\_\_\_\_\_ < x < \_\_\_\_\_\_\_\_\_\_\_) = \_\_\_\_\_\_\_\_\_\_\_$

Exercise:

Problem:

Find the 70^{th} percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.



b. $P(x < k)$ = __________ Therefore, k = _________

Solution:

a. Check student’s solution.
b. 0.70, 4.78 years

Homework

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Exercise:

Problem:

What is the probability of spending more than two days in recovery?

a. 0.0580
b. 0.8447
c. 0.0553
d. 0.9420

Exercise:

Problem: The 90^{th} percentile for recovery times is?

a. 8.89
b. 7.07
c. 7.99
d. 4.32

Solution:

c

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

Exercise:

Problem:

Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?

a. Yes
b. No
c. Unable to determine

Exercise:

Problem:

Find the probability that it takes at least eight minutes to find a parking space.

a. 0.0001
b. 0.9270

c. 0.1862
d. 0.0668

Solution:

d

Exercise:

Problem:

Seventy percent of the time, it takes more than how many minutes to find a parking space?

a. 1.24
b. 2.41
c. 3.95
d. 6.05

Exercise:

Problem:

According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X = height of the individual.

a. X ~ _____(_____,_____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
c. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.

Solution:

a. $X \sim N(66, 2.5)$

b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082

Exercise:

Problem:

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let *X* = IQ of an individual.

a. *X* ~ _____(_____,_____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
d. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

Exercise:

Problem:

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let *X* = percent of fat calories.

a. *X* ~ _____(_____,_____)
b. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
c. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

Solution:

a. $X \sim N(36, 10)$
b. The probability that a person consumes more than 40% of their calories as fat is 0.3446.
c. Approximately 25% of people consume less than 29.26% of their calories as fat.

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If X = distance in feet for a fly ball, then $X \sim$ ______(______,______)
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

Exercise:

Problem:

In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

a. In words, define the random variable X.
b. $X \sim$ ______(______,______)

c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percent of the children spend over ten hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

Solution:

a. X = number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. $X \sim N(3, 1.5)$
c. The probability that the child spends less than one hour a day unsupervised is 0.0918.
d. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
e. 2.21 hours

Exercise:

Problem:

In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let X = number of votes for President Clinton for an election district.

a. State the approximate distribution of X.
b. Is 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.

d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
e. Find the third quartile for votes for President Clinton.

Exercise:

Problem:

Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

a. In words, define the random variable *X*.
b. *X* ~ _____(_____,_____)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

Solution:

a. *X* = the distribution of the number of days a particular type of criminal trial will take
b. *X* ~ *N*(21, 7)
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

Exercise:

Problem:

Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

a. In words, define the random variable *X*.
b. *X* ~ ______(______,______)
c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3% of her laps are under ______.
e. The middle 80% of her laps are from ________ seconds to ________ seconds.

Exercise:

Problem:

Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let *X* = time in line. [link] displays the ordered real data (in minutes):

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5

3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by X ~ _____(_____,_____)
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren't the answers to part f and part g exactly the same?
i. Why are the answers to part f and part g as close as they are?
j. If only ten customers has been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

Solution:

a. mean = 5.51, s = 2.15
b. Check student's solution.
c. Check student's solution.
d. Check student's solution.
e. $X \sim N(5.51, 2.15)$
f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.

j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit normal curve as well.

Exercise:

Problem:

Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

a. Ricardo's actual GPA is lower than Anita's actual GPA.
b. Ricardo is not passing because his *z*-score is zero.
c. Anita is in the 70^{th} percentile of students at her college.

Exercise:

Problem:

[link] shows a sample of the maximum capacity (maximum number of spectators) of sports stadiums. The table does not include horse-racing or motor-racing stadiums.

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100
51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000

55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107
70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

a. Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums (the data).
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____).
f. Use the distribution in part e to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.
g. Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint: Order the data and count the sports stadiums that have a maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
h. Why aren't the answers to part f and part g exactly the same?

Solution:

1. mean = 60,136
   s = 10,468
2. Answers will vary.

3. Answers will vary.
4. Answers will vary.
5. $X \sim N(60136, 10468)$
6. 0.7440
7. The cumulative relative frequency is 43/60 = 0.717.
8. The answers for part f and part g are not the same, because the normal distribution is only an approximation.

Exercise:

Problem:

An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the *z*-scores first, and then use those to calculate the probability.

Exercise:

Problem:

A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

Solution:

- $n = 100$; $p = 0.1$; $q = 0.9$

- $\mu = np = (100)(0.10) = 10$
- $\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

i. $z = \pm 1$: $x_1 = \mu + z\sigma = 10 + 1(3) = 13$ and $x2 = \mu - z\sigma = 10 - 1(3) = 7$. 68% of the defective cars will fall between seven and 13.
ii. $z = \pm 2$: $x_1 = \mu + z\sigma = 10 + 2(3) = 16$ and $x2 = \mu - z\sigma = 10 - 2(3) = 4$. 95 % of the defective cars will fall between four and 16
iii. $z = \pm 3$: $x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x2 = \mu - z\sigma = 10 - 3(3) = 1$. 99.7% of the defective cars will fall between one and 19.

Exercise:

Problem:

We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

a. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

Exercise:

Problem:

A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

a. somewhere between 34 and 54 prizes.
b. somewhere between 54 and 64 prizes.
c. more than 64 prizes.

Solution:

- n = 190; p = $\frac{1}{5}$ = 0.2; q = 0.8
- μ = np = (190)(0.2) = 38
- σ = $\sqrt{npq}$ = $\sqrt{(190)(0.2)(0.8)}$ = 5.5136

a. For this problem: $P(34 < x < 54)$ = normalcdf(34,54,48,5.5136) = 0.7641
b. For this problem: $P(54 < x < 64)$ = normalcdf(54,64,48,5.5136) = 0.0018
c. For this problem: $P(x > 64)$ = normalcdf(64,10^{99},48,5.5136) = 0.0000012 (approximately 0)

Exercise:

Problem:

Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site.

On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

a. Find the probability that the percent of 18 to 34-year-olds who check Facebook before getting out of bed in the morning is at least 30.
b. Find the 95^{th} percentile, and express it in a sentence.

Introduction class="introduction"

If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**.

The **central limit theorem** (clt for short) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size *n* from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size *n* with a "large enough *n*," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size *n* that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

The size of the sample, *n*, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.**

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data, even if its distribution is not normal, behaves in a predictable way is a powerful tool.

Note:
Collaborative Classroom Activity
Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.
Each time a person rolls more than one die, he or she calculates the sample **mean** of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.
The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.
Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Since the "mean" when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?
Draw the graph for the means using two dice. Do the sample means show any kind of pattern?
Draw the graph for the means using five dice. Do you see any pattern emerging?
Finally, draw the graph for the means using ten dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?
As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt).
The central limit theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

Glossary

Sampling Distribution
: Given simple random samples of size *n* from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable X which consists of sample means, tends to be **normally distributed** and

$X \sim N\left(\mu_x, \frac{\sigma x}{\sqrt{n}}\right)$.

The **central limit theorem** for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and **calculating their means,** the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. Standard deviation is the square root of variance, so the standard deviation of the sampling distribution is the standard deviation of the original distribution divided by the square root of n. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n, the distribution of the random variable X, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as n, the **sample size**, increases.

The random variable X has a different z-score associated with it from that of the random variable X. The mean x is the value of X in one sample.
Equation:

$$z = \frac{x - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

μ_X is the average of both X and X.

$\sigma x = \frac{\sigma x}{\sqrt{n}}$ = standard deviation of X and is called the **standard error of the mean.**

Note:
To find probabilities for means on the calculator, follow these steps.
2nd DISTR
2:normalcdf

$$normalcdf\left(lower\ value\ of\ the\ area,\ upper\ value\ of\ the\ area,\ mean,\ \frac{standard\ deviation}{\sqrt{sample\ size}}\right)$$

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = *n*

Example:
An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size *n* = 25 are drawn randomly from the population.
Exercise:

Problem: a. Find the probability that the **sample mean** is between 85 and 92.

Solution:

a. Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

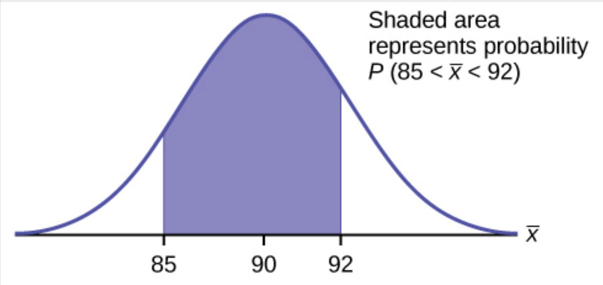
Let X = the mean of a sample of size 25. Since μ_X = 90, σ_X = 15, and n = 25,

$X \sim N\left(90, \frac{15}{\sqrt{25}}\right)$.

Find $P(85 < x < 92)$. Draw a graph.

$P(85 < x < 92) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.



Note:
`normalcdf`(lower value, upper value, mean, standard error of the mean)
The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf`$(85,92,90,\frac{15}{\sqrt{25}})$ = 0.6997

Exercise:

Problem:

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Solution:

b. To find the value that is two standard deviations above the expected value 90, use the formula:

value = μ_x + (#ofTSDEVs)$\left(\frac{\sigma_x}{\sqrt{n}}\right)$

value = 90 + 2 $\left(\frac{15}{\sqrt{25}}\right)$ = 96

The value that is two standard deviations above the expected value is 96.

The standard error of the mean is $\frac{\sigma x}{\sqrt{n}}$ = $\frac{15}{\sqrt{25}}$ = 3. Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size *n*.

Note:
Try It
Exercise:

Problem:

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size *n* = 30 are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Solution:

$P(42 < x < 50)$ = $\left(42,50,45,\frac{8}{\sqrt{30}}\right)$ = 0.9797

Example:
Exercise:

Problem:

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size *n* = 50** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution:

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let $\bar{X}$ = the **mean** time, in hours, it takes to play one soccer match.

If $\mu_{\bar{X}}$ = __________, $\sigma_{\bar{X}}$ = ___________, and n = ____________, then $\bar{X}$ ~ N(_______, _______) by the **central limit theorem for means**.

$\mu_{\bar{X}} = 2$, $\sigma_{\bar{X}} = 0.5$, $n = 50$, and $\bar{X} \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \bar{x} < 2.3)$. Draw a graph.

$P(1.8 < \bar{x} < 2.3) = 0.9977$

`normalcdf`$\left(1.8,2.3,2,\frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Note:
Try It
Exercise:

Problem:

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of n = 60 is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

Solution:

$P(2 < \bar{x} < 3) = \text{normalcdf}\left(2, 3, 2.5, \frac{0.25}{\sqrt{60}}\right) = 1$

Note:
To find percentiles for means on the calculator, follow these steps.
2^{nd} DIStR
3:invNorm
k = invNorm$\left(\text{area to the left of } k, \text{ mean}, \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}\right)$
where:

- k = the k^{th} percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example:
Exercise:

Problem:

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size n = 100.

a. What are the mean and standard deviation for the sample mean ages of tablet users?
b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
d. Find the 95^{th} percentile for the sample mean age (to one decimal place).

Solution:

a. Since the sample mean tends to target the population mean, we have $\mu_\chi = \mu = 34$. The sample standard deviation is given by $\sigma_\chi = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$
b. The central limit theorem states that for large sample sizes(n), the sampling distribution will be approximately normal.
c. The probability that the sample mean age is more than 30 is given by $P(X > 30)$ = `normalcdf`(30,E99,34,1.5) = 0.9962
d. Let k = the 95^{th} percentile.
k = invNorm$\left(0.95,34,\frac{15}{\sqrt{100}}\right)$ = 36.5

Note:
Try It

Exercise:

Problem:

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Solution:

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

$P(29 < x < 35)$ = `normalcdf` $\left(29,35,28,\frac{4.8}{\sqrt{100}}\right)$ = 0.0186

You can conclude there is approximately a 1.9% chance that your game will be played by men whose mean age is between 29 and 35.

Example:
Exercise:

Problem:

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

a. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
b. What is the standard error of the mean?
c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution:

a. $\mu_x = \mu = 8.2$ $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$
b. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.
c. Let k = the 90th percentile
   k = `invNorm` $\left(0.90,8.2,\frac{1}{\sqrt{60}}\right)$ = 8.37. This values indicates that 90 percent of the

average app engagement time for table users is less than 8.37 minutes.

d. $P(8 < x < 8.5)$ = `normalcdf` $\left(8,8.5,8.2,\frac{1}{\sqrt{60}}\right)$ = 0.9293

Note:
Try It
Exercise:

Problem:

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are n = 34, x = 16.01 ounces. If the cans are filled so that μ = 16.00 ounces (as labeled) and σ = 0.143 ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Solution:

We have $P((x > 16.01)$ = `normalcdf` $\left(16.01,\text{E}99,16,\frac{0.143}{\sqrt{34}}\right)$ = 0.3417. Since there is a 34.17% probability that the average sample weight is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

References

Baran, Daya. "20 Percent of Americans Have Never Used Email."WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013).

Data from the United States Department of Agriculture.

Chapter Review

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

Formula Review

The Central Limit Theorem for Sample Means: $X \sim N\left(\mu_x, \frac{\sigma x}{\sqrt{n}}\right)$

The Mean X: μ_x

Central Limit Theorem for Sample Means z-score and standard error of the mean: $z = \frac{x - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$

Standard Error of the Mean (Standard Deviation (X)): $\frac{\sigma_x}{\sqrt{n}}$

Use the following information to answer the next six exercises: Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let X be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

Exercise:

Problem: What is the mean, standard deviation, and sample size?

Solution:

mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

Exercise:

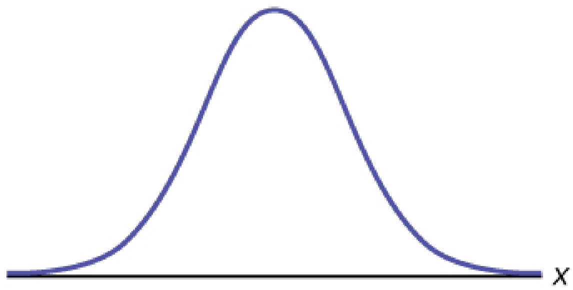
Problem: Complete the distributions.

a. X ~ _____(_____,_____)
b. X ~ _____(_____,_____)

Exercise:

Problem:

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



a.
b. $P(\_\_\_\_\_\_\_\_ < x < \_\_\_\_\_\_\_\_) = \_\_\_\_\_\_\_$

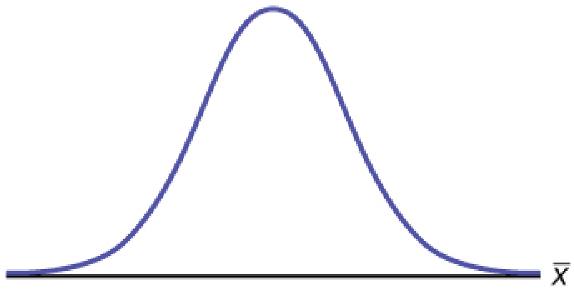
Solution:

a. Check student's solution.
b. 3.5, 4.25, 0.2441

Exercise:

Problem:

Find the probability that the **mean** of a month’s reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



a.
b. $P(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_) = \_\_\_\_\_\_\_$

Exercise:

Problem: What causes the probabilities in [link] and [link] to be different?

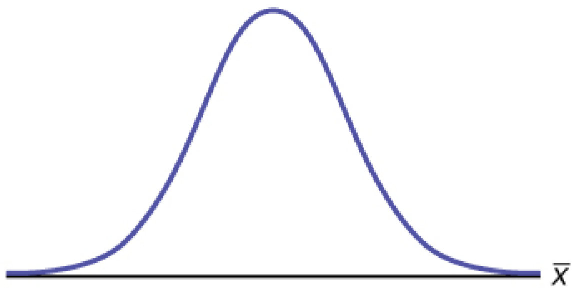
Solution:

The fact that the two distributions are different accounts for the different probabilities.

Exercise:

Problem:

Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.



a.
b. The 95th Percentile =_____________

Homework

Exercise:

Problem:

Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of \$0.88. Suppose that we randomly pick 25 daytime statistics students.

a. In words, X = _____________
b. X ~ _____(_____,_____)
c. In words, X = _____________
d. X ~ ______ (______, ______)
e. Find the probability that an individual had between \$0.80 and \$1.00. Graph the situation, and shade in the area to be determined.
f. Find the probability that the average of the 25 students was between \$0.80 and \$1.00. Graph the situation, and shade in the area to be determined.
g. Explain why there is a difference in part e and part f.

Solution:

a. X = amount of change students carry
b. X ~ E(0.88, 0.88)
c. X = average amount of change carried by a sample of 25 sstudents.
d. X ~ N(0.88, 0.176)
e. 0.0819
f. 0.1882
g. The distributions are different. Part a is exponential and part b is normal.

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If X = average distance in feet for 49 fly balls, then X ~ _______(_______,_______)
b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80^{th} percentile of the distribution of the average of 49 fly balls.

Exercise:

Problem:

According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. In words, X = _____________
b. In words, X = _____________
c. X ~ _____(_____,_____)
d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
e. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

Solution:

a. length of time for an individual to complete IRS form 1040, in hours.
b. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
c. $N\left(10.53,\ \frac{1}{3}\right)$
d. Yes. I would be surprised, because the probability is almost 0.
e. No. I would not be totally surprised because the probability is 0.2312

Exercise:

Problem:

Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let X the average of the 49 races.

a. X ~ _____(_____,_____)
b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
c. Find the 80th percentile for the average of these 49 marathons.
d. Find the median of the average running times.

Exercise:

Problem:

The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

a. In words, X = _________
b. X ~ _____________
c. In words, X = _____________
d. X ~ _____(_____,_____)
e. Find the first quartile for the average song length, X.
f. The IQR (interquartile range) for the average song length, X, is from ___ - ___.

Solution:

a. the length of a song, in minutes, in the collection
b. $U(2, 3.5)$
c. the average length, in minutes, of the songs from a sample of five albums from the collection
d. $N(2.75, 0.0660)$
e. 2.71 minutes
f. 0.09 minutes

Exercise:

Problem:

In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

a. In words, X = _____________
b. In words, X = _____________
c. X ~ _____(_____,_____)
d. The IQR for X is from _______ acres to _______ acres.

Exercise:

Problem:

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of X is approximately equal to the mean of *X*.
b. When the sample size is large, X is approximately normally distributed.
c. When the sample size is large, the standard deviation of X is approximately the same as the standard deviation of *X*.

Solution:

a. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
c. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.

Exercise:

Problem:

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen. Let X = average percent of fat calories.

a. X ~ ______(______, ______)
b. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percent of fat calories.

Exercise:

Problem:

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.

a. In words, *X* = ______________
b. In words, X = ______________

c. X ~ ______(______,______)
d. How is it possible for the standard deviation to be greater than the average?
e. Why is it more likely that the average of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to $2,200?

Solution:

a. X = the yearly income of someone in a third world country
b. the average salary from samples of 1,000 residents of a third world country
c. $\bar{X} \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$
d. Very wide differences in data values can have averages smaller than standard deviations.
e. The distribution of the sample mean will have higher probabilities closer to the population mean.
$P(2000 < \bar{X} < 2100) = 0.1537$
$P(2100 < \bar{X} < 2200) = 0.1317$

Exercise:

Problem: Which of the following is NOT TRUE about the distribution for averages?

a. The mean, median, and mode are equal.
b. The area under the curve is one.
c. The curve never touches the *x*-axis.
d. The curve is skewed to the right.

Exercise:

Problem:

The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations. The distribution to use for the average cost of gasoline for the 16 gas stations is:

a. $\bar{X} \sim N(4.59, 0.10)$
b. $\bar{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
c. $\bar{X} \sim N\left(4.59, \frac{16}{0.10}\right)$
d. $\bar{X} \sim N\left(4.59, \frac{\sqrt{16}}{0.10}\right)$

Solution:

b

Glossary

Average
: a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem
: Given a random variable (RV) with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, X, and the sample sum, ΣX. If the size (n) of the sample is sufficiently large, then $X \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ and $\Sigma X \sim N(n\mu, (\sqrt{n})(\sigma))$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean, and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Normal Distribution
: a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called a **standard normal distribution**.

Standard Error of the Mean
: the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

Using the Central Limit Theorem

It is important for you to understand when to use the **central limit theorem**. If you are being asked to find the probability of the mean, use the clt for the mean. If you are being asked to find the probability of a sum or total, use the clt for sums. This also applies to percentiles for means and sums.

Note:
NOTE
If you are being asked to find the probability of an **individual** value, do **not** use the clt. **Use the distribution of its random variable.**

Examples of the Central Limit Theorem

Law of Large Numbers

The **law of large numbers** says that if you take samples of larger and larger size from any population, then the mean x of the sample tends to get closer and closer to μ. From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for X is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean x must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

Central Limit Theorem for the Mean and Sum Examples

Example:
A study involving stress is conducted among the students on a college campus. **The stress scores follow a uniform distribution** with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:

a. The probability that the **mean stress score** for the 75 students is less than two.
b. The 90^{th} percentile for the **mean stress score** for the 75 students.
c. The probability that the **total of the 75 stress scores** is less than 200.
d. The 90^{th} percentile for the **total stress score** for the 75 students.

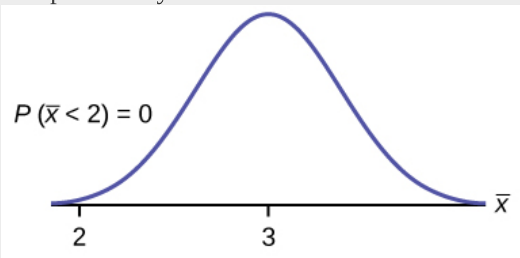
Let X = one stress score.
Problems a and b ask you to find a probability or a percentile for a **mean**. Problems c and d ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.
Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$ (See Continuous Random Variables for an explanation on the uniform distribution).
$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$
$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$
For problems a. and b., let X = the mean stress score for the 75 students. Then,
$X \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$
Exercise:

Problem: a. Find $P(x < 2)$. Draw the graph.

Solution:

a. $P(x < 2) = 0$

The probability that the mean stress score is less than two is about zero.



normalcdf $\left(1,2,3,\frac{1.15}{\sqrt{75}}\right) = 0$

> **Note:**
> **Reminder**
> The smallest stress score is one.

Exercise:

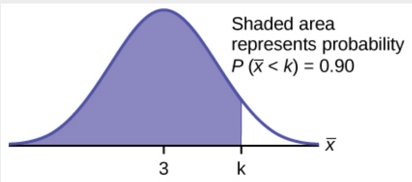
Problem: b. Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

Solution:

b. Let k = the 90th precentile.

Find k, where $P(x < k) = 0.90$.

$k = 3.2$



The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2, and that 10% are at least 3.2.

invNorm$\left(0.90,3,\frac{1.15}{\sqrt{75}}\right) = 3.2$

For problems c and d, let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N[(75)(3),(\sqrt{75})(1.15)]$

Exercise:

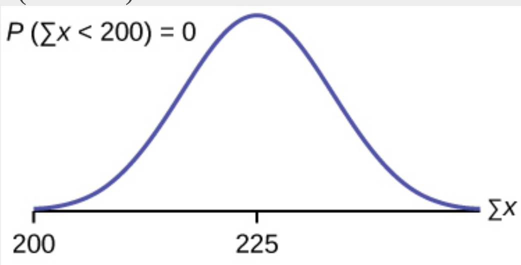
Problem: c. Find $P(\Sigma x < 200)$. Draw the graph.

Solution:

c. The mean of the sum of 75 stress scores is (75)(3) = 225

The standard deviation of the sum of 75 stress scores is $(\sqrt{75})(1.15) = 9.96$

$P(\Sigma x < 200) = 0$



The probability that the total of 75 scores is less than 200 is about zero.

`normalcdf` (75,200,(75)(3),$(\sqrt{75})$(1.15)).

Note:
Reminder
The smallest total of 75 stress scores is 75, because the smallest single score is one.

Exercise:

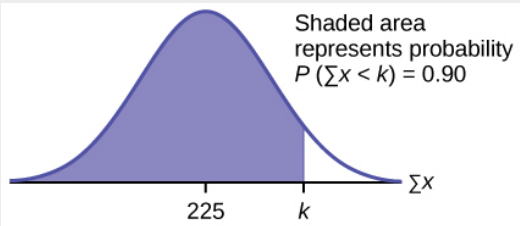
Problem: d. Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution:

d. Let k = the 90th percentile.

Find k where $P(\Sigma x < k) = 0.90$.

$k = 237.8$



The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

`invNorm`(0.90,(75)(3),$(\sqrt{75})$(1.15)) = 237.8

Note:
Try It
Exercise:

Problem: Use the information in [link], but use a sample size of 55 to answer the following questions.

a. Find $P(x < 7)$.
b. Find $P(\Sigma x > 170)$.
c. Find the 80th percentile for the mean of 55 scores.
d. Find the 85th percentile for the sum of 55 scores.

Solution:
Solutions

a. 0.0265
b. 0.2789
c. 3.13
d. 173.84

Example:
Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes.
Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.
Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.
$X \sim Exp\left(\frac{1}{22}\right)$. From previous chapters, we know that $\mu = 22$ and $\sigma = 22$.
Let X = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.
$X \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the central limit theorem for sample means
Exercise:

Problem:
Using the clt to find probability

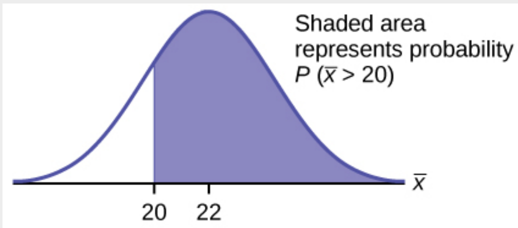
a. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(x > 20)$. Draw the graph.
b. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$.
c. Explain why the probabilities in parts a and b are different.

Solution:

a. Find: $P(x > 20)$

$P(x > 20) = 0.79199$ using `normalcdf` $\left(20{,}1E99{,}22{,}\frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



Note:
Reminder
1E99 = 10^{99} and –1E99 = –10^{99}. Press the `EE` key for E. Or just use 10^{99} instead of 1E99.

b. Find P(x > 20). Remember to use the exponential distribution for an **individual:** $X \sim Exp\left(\frac{1}{22}\right)$.
$P(x > 20) = e^{\left(-\left(\frac{1}{22}\right)(20)\right)}$ or $e^{(-0.04545(20))} = 0.4029$

c.
1. $P(x > 20) = 0.4029$ but $P(x > 20) = 0.7919$
2. The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.
3. When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the clt. Use the clt with the normal distribution when you are being asked to find the probability for a mean.

Exercise:

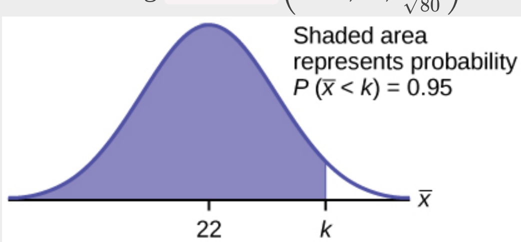
Problem:

Using the clt to find percentiles
Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution:

Let k = the 95th percentile. Find k where $P(x < k) = 0.95$

$k = 26.0$ using `invNorm` $\left(0.95{,}22{,}\frac{22}{\sqrt{80}}\right) = 26.0$



The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

Ninety five percent of such samples would have means under 26 minutes; only five percent of such samples would have means above 26 minutes.

Note:
Try It
Exercise:

Problem: Use the information in [link], but change the sample size to 144.

a. Find $P(20 < \bar{x} < 30)$.
b. Find $P(\Sigma x$ is at least 3,000).
c. Find the 75th percentile for the sample mean excess time of 144 customers.
d. Find the 85th percentile for the sum of 144 excess times used by customers.

Solution:
Solutions

a. 0.8623
b. 0.7377
c. 23.2
d. 3,441.6

Example:
In the United States, someone is sexually assaulted every two minutes, on average, according to a number of studies. Suppose the standard deviation is 0.5 minutes and the sample size is 100.
Exercise:

Problem:

a. Find the median, the first quartile, and the third quartile for the sample mean time of sexual assaults in the United States.
b. Find the median, the first quartile, and the third quartile for the sum of sample times of sexual assaults in the United States.
c. Find the probability that a sexual assault occurs on the average between 1.75 and 1.85 minutes.
d. Find the value that is two standard deviations above the sample mean.
e. Find the *IQR* for the sum of the sample times.

Solution:

a. We have, $\mu_x = \mu = 2$ and $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore:

   1. 50th percentile = $\mu_x = \mu = 2$
   2. 25th percentile = *invNorm*(0.25,2,0.05) = 1.97
   3. 75th percentile = *invNorm*(0.75,2,0.05) = 2.03

b. We have $\mu_{\Sigma x} = n(\mu_x) = 100(2) = 200$ and $\sigma_{\mu x} = \sqrt{n}(\sigma_x) = 10(0.5) = 5$. Therefore

1. 50th percentile = $\mu_{\Sigma x} = n(\mu_x) = 100(2) = 200$
2. 25th percentile = invNorm(0.25,200,5) = 196.63
3. 75th percentile = invNorm(0.75,200,5) = 203.37

c. $P(1.75 < x < 1.85)$ = `normalcdf`(1.75,1.85,2,0.05) = 0.0013
d. Using the z-score equation, $z = \frac{x-\mu_x}{\sigma_x}$, and solving for x, we have $x = 2(0.05) + 2 = 2.1$
e. The *IQR* is 75th percentile – 25th percentile = 203.37 – 196.63 = 6.74

Note:
Try It
Exercise:

Problem:

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mm Hg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follow a normal distribution.

a. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
b. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.
c. If the sample were four women between the ages of 18 to 24 and we did not know the original distribution, could the central limit theorem be used?

Solution:

a. $P(x > 120)$ = `normalcdf`(120,1E99,114.8,13.1) = 0.3457. There is about a 35% chance, that the randomly selected woman will have systolics blood pressure greater than 120.
b. $P(x > 120)$ = `normalcdf`$\left(120,\ 1\text{E}99{,}114.8{,}\frac{13.1}{\sqrt{40}}\right)$ = 0.006. There is only a 0.6% chance that the average systolic blood pressure for the randomly selected group is greater than 120.
c. The central limit theorem could not be used if the sample size were four and we did not know the original distribution was normal. The sample size would be too small.

Example:
Exercise:

Problem:

A study was done about violence against prostitutes and the symptoms of the posttraumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

a. In a sample of 25 prostitutes, what is the probability that the mean age of the prostitutes is less than 35?
b. Is it likely that the mean age of the sample group could be more than 50 years? Interpret the results.

c. In a sample of 49 prostitutes, what is the probability that the sum of the ages is no less than 1,600?
d. Is it likely that the sum of the ages of the 49 prostitutes is at most 1,595? Interpret the results.
e. Find the 95th percentile for the sample mean age of 65 prostitutes. Interpret the results.
f. Find the 90th percentile for the sum of the ages of 65 prostitutes. Interpret the results.

Solution:

a. $P(\bar{x} < 35)$ = `normalcdf`(-E99,35,30.9,1.8) = 0.9886
b. $P(\bar{x} > 50)$ = `normalcdf`(50, E99,30.9,1.8) ≈ 0. For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.
c. $P(\Sigma x \geq 1{,}600)$ = `normalcdf`(1600,E99,1514.10,63) = 0.0864
d. $P(\Sigma x \leq 1{,}595)$ = `normalcdf`(-E99,1595,1514.10,63) = 0.9005. This means that there is a 90% chance that the sum of the ages for the sample group n = 49 is at most 1595.
e. The 95th percentile = `invNorm`(0.95,30.9,1.1) = 32.7. This indicates that 95% of the prostitutes in the sample of 65 are younger than 32.7 years, on average.
f. The 90th percentile = `invNorm`(0.90,2008.5,72.56) = 2101.5. This indicates that 90% of the prostitutes in the sample of 65 have a sum of ages less than 2,101.5 years.

Note:
Try It
Exercise:

Problem:

According to Boeing data, the 757 airliner carries 200 passengers and has doors with a height of 72 inches. Assume for a certain population of men we have a mean height of 69.0 inches and a standard deviation of 2.8 inches.

a. What doorway height would allow 95% of men to enter the aircraft without bending?
b. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
c. For engineers designing the 757, which result is more relevant: the height from part a or part b? Why?

Solution:

a. We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 2.8$. The height of the doorway is found to be `invNorm`(0.95,69,2.8) = 73.61
b. We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 0.28$. So, `invNorm`(0.95,69,0.28) = 69.49
c. When designing the doorway heights, we need to incorporate as much variability as possible in order to accommodate as many passengers as possible. Therefore, we need to use the result based on part a.

Note:
HISTORICAL NOTE
: Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n(say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n, you had to use the binomial formula, which could be very complicated. Using the **normal approximation to the binomial** distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 – p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x – 0.5$). The number 0.5 is called the **continuity correction factor** and is used in the following example.

Example:
Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

a. Find the probability that **at least 150** favor a charter school.
b. Find the probability that **at most 160** favor a charter school.
c. Find the probability that **more than 155** favor a charter school.
d. Find the probability that **fewer than 147** favor a charter school.
e. Find the probability that **exactly 175** favor a charter school.

Let X = the number that favor a charter school for grades K trough 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y. $Y \sim N(159, 8.6447)$. See The Normal Distribution for help with calculator instructions.

For part a, you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.
`normalcdf`(149.5,10^99,159,8.6447) = 0.8641.
For part b, you **include 160** so $P(X \leq 160)$ has normal appraximation $P(Y \leq 160.5) = 0.5689$.
`normalcdf`(0,160.5,159,8.6447) = 0.5689
For part c, you **exclude 155** so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$.
`normalcdf`(155.5,10^99,159,8.6447) = 0.6572.
For part d, you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.
`normalcdf`(0,146.5,159,8.6447) = 0.0741
For part e,$P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.
`normalcdf`(174.5,175.5,159,8.6447) = 0.0083
Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators, and they easily calculate probabilities for the binomial distribution. If you type in "binomial probability distribution calculation" in an Internet browser, you can find at least one online calculator for the binomial. For [link], the probabilities are calculated using the following binomial distribution: ($n = 300$ and $p = 0.53$). Compare the binomial and normal distribution answers. See Discrete Random Variables for help with calculator instructions for the binomial.

$P(X \geq 150)$:`1 - binomialcdf`(300,0.53,149) = 0.8641
$P(X \leq 160)$:`binomialcdf`(300,0.53,160) = 0.5684
$P(X > 155)$:`1 - binomialcdf`(300,0.53,155) = 0.6576
$P(X < 147)$:`binomialcdf`(300,0.53,146) = 0.0742
$P(X = 175)$:(You use the binomial pdf.)`binomialpdf`(300,0.53,175) = 0.0083

Note:
Try It
Exercise:

Problem:

In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

Solution:
Solutions

0.0401

References

Data from the Wall Street Journal.

"National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).

Chapter Review

The central limit theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer the sample mean x gets to μ.

Use the following information to answer the next ten exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is taken.
Exercise:

Problem:

a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deivation?
b. What is the distribution for the mean weight of 100 25-pound lifting weights?
c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

Solution:

a. $U(24, 26)$, 25, 0.5774
b. $N(25, 0.0577)$
c. 0.0416

Exercise:

Problem: Draw the graph from [link]

Exercise:

Problem: Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

Solution:

0.0003

Exercise:

Problem: Draw the graph from [link]

Exercise:

Problem: Find the 90th percentile for the mean weight for the 100 weights.

Solution:

25.07

Exercise:

Problem: Draw the graph from [link]

Exercise:

Problem:

a. What is the distribution for the sum of the weights of 100 25-pound lifting weights?
b. Find $P(\Sigma x < 2{,}450)$.

Solution:

a. $N(2{,}500, 5.7735)$
b. 0

Exercise:

Problem: Draw the graph from [link]

Exercise:

Problem: Find the 90th percentile for the total weight of the 100 weights.

Solution:

2,507.40

Exercise:

Problem: Draw the graph from [link]

Use the following information to answer the next five exercises: The length of time a particular smartphone's battery lasts follows an exponential distribution with a mean of ten months. A sample of 64 of these smartphones is taken.

Exercise:

Problem:

a. What is the standard deviation?
b. What is the parameter *m*?

Solution:

a. 10
b. $\frac{1}{10}$

Exercise:

Problem: What is the distribution for the length of time one battery lasts?

Exercise:

Problem: What is the distribution for the mean length of time 64 batteries last?

Solution:

$N\left(10, \frac{10}{8}\right)$

Exercise:

Problem: What is the distribution for the total length of time 64 batteries last?

Exercise:

Problem: Find the probability that the sample mean is between seven and 11.

Solution:

0.7799

Exercise:

Problem: Find the 80th percentile for the total length of time 64 batteries last.

Exercise:

Problem: Find the *IQR* for the mean amount of time 64 batteries last.

Solution:

1.69

Exercise:

Problem: Find the middle 80% for the total amount of time 64 batteries last.

Use the following information to answer the next eight exercises: A uniform distribution has a minimum of six and a maximum of ten. A sample of 50 is taken.

Exercise:

Problem: Find $P(\Sigma x > 420)$.

Solution:

0.0072

Exercise:

Problem: Find the 90th percentile for the sums.

Exercise:

Problem: Find the 15th percentile for the sums.

Solution:

391.54

Exercise:

Problem: Find the first quartile for the sums.

Exercise:

Problem: Find the third quartile for the sums.

Solution:

405.51

Exercise:

Problem: Find the 80th percentile for the sums.

Homework

Exercise:

Problem:

The attention span of a two-year-old is exponentially distributed with a mean of about eight minutes. Suppose we randomly survey 60 two-year-olds.

a. In words, X = ________
b. X ~ ______(______,______)
c. In words, $\bar{X}$ = ______________
d. $\bar{X}$ ~ ______(______,______)
e. Before doing any calculations, which do you think will be higher? Explain why.

  i. The probability that an individual attention span is less than ten minutes.
  ii. The probability that the average attention span for the 60 children is less than ten minutes?

f. Calculate the probabilities in part e.
g. Explain why the distribution for X is not exponential.

Exercise:

Problem: The closing stock prices of 35 U.S. semiconductor manufacturers are given as follows.

8.625 30.25 27.625 46.75 32.875 18.25 5 0.125 2.9375 6.875 28.25 24.25 21 1.5 30.25 71 43.5 49.25 2.5625 31 16.5 9.5 18.5 18 9 10.5 16.625 1.25 18 12.87 7 12.875 2.875 60.25 29.25

a. In words, X = _______________

b.
i. x = ______
ii. s_x = ______
iii. n = ______

c. Construct a histogram of the distribution of the averages. Start at x = –0.0005. Use bar widths of ten.
d. In words, describe the distribution of stock prices.
e. Randomly average five stock prices together. (Use a random number generator.) Continue averaging five pieces together until you have ten averages. List those ten averages.
f. Use the ten averages from part e to calculate the following.

i. x = ______
ii. s_x = ______

g. Construct a histogram of the distribution of the averages. Start at x = -0.0005. Use bar widths of ten.
h. Does this histogram look like the graph in part c?
i. In one or two complete sentences, explain why the graphs either look the same or look different?
j. Based upon the theory of the **central limit theorem**, X ~ _____(_____,____)

Solution:

a. X = the closing stock prices for U.S. semiconductor manufacturers
b. i. \$20.71; ii. \$17.31; iii. 35
c.
d. Exponential distribution, $X \sim Exp\left(\frac{1}{20.71}\right)$
e. Answers will vary.
f. i. \$20.71; ii. \$11.14
g. Answers will vary.
h. Answers will vary.
i. Answers will vary.
j. $N\left(20.71,\ \frac{17.31}{\sqrt{5}}\right)$

Use the following information to answer the next three exercises: Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

Exercise:

Problem: X ~ _____(_____,_____)

a. $U(0,4)$
b. $U(10,2)$
c. $E\chi p(2)$
d. $N(2,1)$

Exercise:

Problem: The average wait time is:

a. one hour.
b. two hours.
c. two and a half hours.
d. four hours.

Solution:

b

Exercise:

Problem:

Suppose that it is now past noon on a delivery day. The probability that a person must wait at least one and a half **more** hours is:

a. $\frac{1}{4}$
b. $\frac{1}{2}$
c. $\frac{3}{4}$
d. $\frac{3}{8}$

Use the following information to answer the next two exercises: The time to wait for a particular rural bus is distributed uniformly from zero to 75 minutes. One hundred riders are randomly sampled to learn how long they waited.

Exercise:

Problem: The 90th percentile sample average wait time (in minutes) for a sample of 100 riders is:

a. 315.0
b. 40.3
c. 38.5
d. 65.2

Solution:

b

Exercise:

Problem:

Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

a. yes
b. no
c. There is not enough information.

Use the following to answer the next two exercises: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

Exercise:

Problem: What's the approximate probability that the average price for 16 gas stations is over $4.69?

a. almost zero
b. 0.1587
c. 0.0943
d. unknown

Solution:

a

Exercise:

Problem: Find the probability that the average price for 30 gas stations is less than $4.55.

a. 0.6554
b. 0.3446
c. 0.0142
d. 0.9858
e. 0

Exercise:

Problem:

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through five. A simple random sample of 300 is surveyed. Calculate following using the normal approximation to the binomial distribtion.

a. Find the probability that less than 100 favor a charter school for grades K through 5.
b. Find the probability that 170 or more favor a charter school for grades K through 5.
c. Find the probability that no more than 140 favor a charter school for grades K through 5.
d. Find the probability that there are fewer than 130 that favor a charter school for grades K through 5.
e. Find the probability that exactly 150 favor a charter school for grades K through 5.

If you have access to an appropriate calculator or computer software, try calculating these probabilities using the technology.

Solution:

a. 0
b. 0.1123

c. 0.0162
d. 0.0003
e. 0.0268

Exercise:

Problem:

Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities. Round the standard deviation to four decimal places.

a. Find the probability that Janice is the driver at most 20 days.
b. Find the probability that Roberta is the driver more than 16 days.
c. Find the probability that Barbara drives exactly 24 of those 96 days.

Exercise:

Problem:

$X \sim N(60, 9)$. Suppose that you form random samples of 25 from this distribution. Let X be the random variable of averages. Let ΣX be the random variable of sums. For parts c through f, sketch the graph, shade the region, label and scale the horizontal axis for X, and find the probability.

a. Sketch the distributions of X and X on the same graph.
b. X ~ ______(______,______)
c. $P(x < 60)$ = ______
d. Find the 30th percentile for the mean.
e. $P(56 < x < 62)$ = ______
f. $P(18 < x < 58)$ = ______
g. Σx ~ ______(______,______)
h. Find the minimum value for the upper quartile for the sum.
i. $P(1{,}400 < \Sigma x < 1{,}550)$ = ______

Solution:

a. Check student's solution.
b. $X \sim N\left(60, \frac{9}{\sqrt{25}}\right)$
c. 0.5000
d. 59.06
e. 0.8536
f. 0.1333
g. $N(1500, 45)$
h. 1530.35
i. 0.6877

Exercise:

Problem:

Suppose that the length of research papers is uniformly distributed from ten to 25 pages. We survey a class in which 55 research papers were turned in to a professor. The 55 research papers are considered a random collection of all papers. We are interested in the average length of the research papers.

a. In words, X = _______________
b. X ~ ______(______,______)
c. μ_X = ______
d. σ_X = ______
e. In words, X = ________________
f. X ~ ______(______,______)
g. In words, ΣX = _______________
h. ΣX ~ ______(______,______)
i. Without doing any calculations, do you think that it's likely that the professor will need to read a total of more than 1,050 pages? Why?
j. Calculate the probability that the professor will need to read a total of more than 1,050 pages.
k. Why is it so unlikely that the average length of the papers will be less than 12 pages?

Exercise:

Problem:

Salaries for teachers in a particular elementary school district are normally distributed with a mean of $44,000 and a standard deviation of $6,500. We randomly survey ten teachers from that district.

a. Find the 90^{th} percentile for an individual teacher's salary.
b. Find the 90^{th} percentile for the average teacher's salary.

Solution:

a. $52,330
b. $46,634

Exercise:

Problem:

The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

a. In words, X = _______________
b. In words, X = ______________________
c. X ~ ______(______,______)
d. In words, ΣX = _________________
e. ΣX ~ ______(______,______)
f. Is it likely that an individual stayed more than five days in the hospital? Why or why not?
g. Is it likely that the average stay for the 80 women was more than five days? Why or why not?
h. Which is more likely:

    i. An individual stayed more than five days.
    ii. the average stay of 80 women was more than five days.

i. If we were to sum up the women's stays, is it likely that, collectively they spent more than a year in the hospital? Why or why not?

For each problem, wherever possible, provide graphs and use the calculator.

Exercise:

Problem:

NeverReady batteries has engineered a newer, longer lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less? Is the company's claim reasonable?

Solution:

- We have μ = 17, σ = 0.8, x = 16.7, and n = 30. To calculate the probability, we use `normalcdf`(lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$) = `normalcdf`$\left(E\text{–}99{,}16.7{,}17{,}\frac{0.8}{\sqrt{30}}\right)$ = 0.0200.
- If the process is working properly, then the probability that a sample of 30 batteries would have at most 16.7 lifetime hours is only 2%. Therefore, the class was justified to question the claim.

Exercise:

Problem: Men have an average weight of 172 pounds with a standard deviation of 29 pounds.

a. Find the probability that 20 randomly selected men will have a sum weight greater than 3600 lbs.
b. If 20 men have a sum weight greater than 3500 lbs, then their total weight exceeds the safety limits for water taxis. Based on (a), is this a safety concern? Explain.

Exercise:

Problem:

M&M candies large candy bags have a claimed net weight of 396.9 g. The standard deviation for the weight of the individual candies is 0.017 g. The following table is from a stats experiment conducted by a statistics class.

Red	Orange	Yellow	Brown	Blue	Green
0.751	0.735	0.883	0.696	0.881	0.925
0.841	0.895	0.769	0.876	0.863	0.914
0.856	0.865	0.859	0.855	0.775	0.881
0.799	0.864	0.784	0.806	0.854	0.865
0.966	0.852	0.824	0.840	0.810	0.865
0.859	0.866	0.858	0.868	0.858	1.015
0.857	0.859	0.848	0.859	0.818	0.876
0.942	0.838	0.851	0.982	0.868	0.809

Red	Orange	Yellow	Brown	Blue	Green
0.873	0.863			0.803	0.865
0.809	0.888			0.932	0.848
0.890	0.925			0.842	0.940
0.878	0.793			0.832	0.833
0.905	0.977			0.807	0.845
	0.850			0.841	0.852
	0.830			0.932	0.778
	0.856			0.833	0.814
	0.842			0.881	0.791
	0.778			0.818	0.810
	0.786			0.864	0.881
	0.853			0.825	
	0.864			0.855	
	0.873			0.942	
	0.880			0.825	
	0.882			0.869	
	0.931			0.912	
				0.887	

The bag contained 465 candies and he listed weights in the table came from randomly selected candies. Count the weights.

a. Find the mean sample weight and the standard deviation of the sample weights of candies in the table.
b. Find the sum of the sample weights in the table and the standard deviation of the sum of the weights.
c. If 465 M&Ms are randomly selected, find the probability that their weights sum to at least 396.9.
d. Is the Mars Company's M&M labeling accurate?

Solution:

a. For the sample, we have $n = 100$, $x = 0.862$, $s = 0.05$
b. $\Sigma x = 85.65$, $\Sigma s = 5.18$
c. `normalcdf`$(396.9, E99, (465)(0.8565), (0.05)(\sqrt{465})) \approx 1$
d. Since the probability of a sample of size 465 having at least a mean sum of 396.9 is appproximately 1, we can conclude that Mars is correctly labeling their M&M packages.

Exercise:

Problem:

The Screw Right Company claims their $\frac{3}{4}$ inch screws are within ±0.23 of the claimed mean diameter of 0.750 inches with a standard deviation of 0.115 inches. The following data were recorded.

0.757	0.723	0.754	0.737	0.757	0.741	0.722	0.741	0.743	0.742
0.740	0.758	0.724	0.739	0.736	0.735	0.760	0.750	0.759	0.754
0.744	0.758	0.765	0.756	0.738	0.742	0.758	0.757	0.724	0.757
0.744	0.738	0.763	0.756	0.760	0.768	0.761	0.742	0.734	0.754
0.758	0.735	0.740	0.743	0.737	0.737	0.725	0.761	0.758	0.756

The screws were randomly selected from the local home repair store.

a. Find the mean diameter and standard deviation for the sample
b. Find the probability that 50 randomly selected screws will be within the stated tolerance levels. Is the company's diameter claim plausible?

Exercise:

Problem:

Your company has a contract to perform preventive maintenance on thousands of air-conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

Solution:

Use `normalcdf`

$$\left(E\text{–}99,1.1,1,\frac{1}{\sqrt{70}}\right)$$

= 0.7986. This means that there is an 80% chance that the service time will be less than 1.1 hours. It could be wise to schedule more time since there is an associated 20% chance that the maintenance time will be greater than 1.1 hours.

Exercise:

Problem:

A typical adult has an average IQ score of 105 with a standard deviation of 20. If 20 randomly selected adults are given an IQ tesst, what is the probability that the sample mean scores will be between 85 and 125 points?

Exercise:

Problem:

Certain coins have an average weight of 5.201 grams with a standard deviation of 0.065 g. If a vending machine is designed to accept coins whose weights range from 5.111 g to 5.291 g, what is the expected number of rejected coins when 280 randomly selected coins are inserted into the machine?

Solution:

We assume that the weights of coins are normally distributed in the population. Since we have `normalcdf`$\left(5.111,5.291,5.201,\frac{0.065}{\sqrt{280}}\right)$ ≈ 0.8338, we expect (1 – 0.8338)280 ≈ 47 coins to be rejected.

Glossary

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital, notation: $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Mean

a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation.; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$; often referred as the **Rectangular Distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Introduction
class="introduction"

Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's *t* distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, x, and the sample standard deviation, *s*. You would use x to estimate the population mean and *s* to estimate the population standard deviation. The sample mean, x, is the **point estimate** for the population mean, *μ*. The sample standard deviation, *s*, is the point estimate for the population standard deviation, *σ*.

Each of x and *s* is called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. It provides a range of reasonable values in which we expect the population parameter to fall. There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success.

Suppose, for the iTunes example, we do not know the population mean *μ*, but we do know that the population standard deviation is *σ* = 1 and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **empirical rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, x, will be within two standard deviations of the population mean *μ*. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean x is likely to be within 0.2 units of *μ*.

Because x is within 0.2 units of *μ*, which is unknown, then *μ* is likely to be within 0.2 units of x in 95% of the samples. The population mean *μ* is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, *μ* is between $x - 0.2$ and $x + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $x = 2$. Then the unknown population mean μ is between

$x - 0.2 = 2 - 0.2 = 1.8$ and $x + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an x that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Note:
Note
Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Note:
Collaborative Exercise
Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval $\left(x - 2\cdot\frac{\sigma}{\sqrt{n}}, x + 2\cdot\frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between __________ and __________.

Glossary

Confidence Interval (CI)
: an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics
: also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Parameter
: a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

A Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $x = 10$ and we have constructed the 90% confidence interval (5, 15) where *EBM* = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean *μ*, **where the population standard deviation is known**, we need x as an estimate for *μ* and we need the margin of error. Here, the margin of error (*EBM*) is called the **error bound for a population mean** (abbreviated ***EBM***). The sample mean x is the **point estimate** of the unknown population mean *μ*.

The confidence interval estimate will have the form:

(point estimate - error bound, point estimate + error bound) or, in symbols,($x - EBM, x + EBM$)

The margin of error (*EBM*) depends on the **confidence level** (abbreviated ***CL***). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (*α*). *α* is related to the confidence level, *CL*. *α* is the probability that the interval does not contain the unknown population parameter.
Mathematically, *α* + *CL* = 1.

Example:

- Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population.

- The sample mean is seven, and the error bound for the mean is 2.5.

x = 7 and *EBM* = 2.5
The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5).
If the confidence level (*CL*) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Note:
Try It
Exercise:

Problem:

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2.

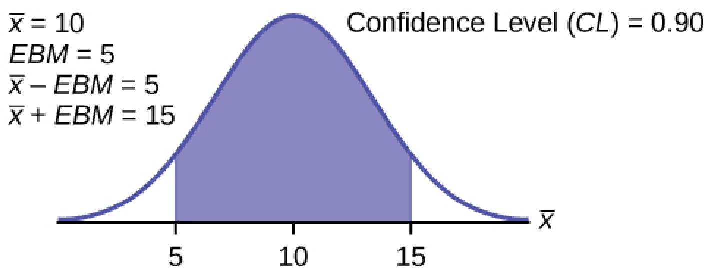
What is the confidence interval estimate for the population mean?

Solution:

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of x = 10, and we have constructed the 90% confidence interval (5, 15) where *EBM* = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.



To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the *z*-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the "standard error of the mean" in order to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- X is normally distributed, that is, $X \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.**

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean x from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the *z*-score that corresponds to the confidence level.
- Calculate the error bound *EBM*.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of

the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the *z*-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 – \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 – 0.025 = 0.975$.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a standard normal probability table.

Note:
`invNorm`(0.975, 0, 1) = 1.96

Note:
Note
Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal

distribution $Z \sim N(0, 1)$.

Calculating the Error Bound (*EBM*)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

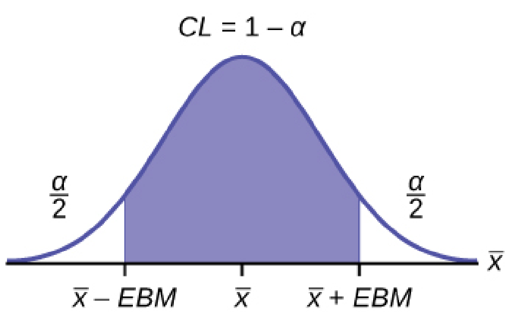
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(x - EBM, x + EBM)$.

The graph gives a picture of the entire situation.

$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$.



Writing the Interpretation

The interpretation should clearly state the confidence level (*CL*), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example:
Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).
Exercise:

Problem:

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution:

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A
To find the confidence interval, you need the sample mean, x, and the *EBM*.

- x = 68
- *EBM* = $\left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; n = 36; The confidence level is 90% (CL = 0.90)

CL = 0.90 so α = 1 – CL = 1 – 0.90 = 0.10

$\frac{\alpha}{2}$ = 0.05 $z_{\frac{\alpha}{2}} = z_{0.05}$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

using invNorm(0.95, 0, 1) on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$

x - *EBM* = 68 - 0.8225 = 67.1775

x + *EBM* = 68 + 0.8225 = 68.8225

The 90% confidence interval is **(67.1775, 68.8225).**

Solution:

Solution B

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `7:ZInterval`.
Press `ENTER`.
Arrow to `Stats` and press `ENTER`.
Arrow down and enter three for σ, 68 for x, 36 for n, and .90 for `C-level`.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation
We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level
Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Note:

Try It
Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Exercise:

Problem:

Find a 90% confidence interval estimate for the population mean delivery time.

Solution:

(34.1347, 37.8653)

Example:
The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. [link] shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Exercise:

Problem:

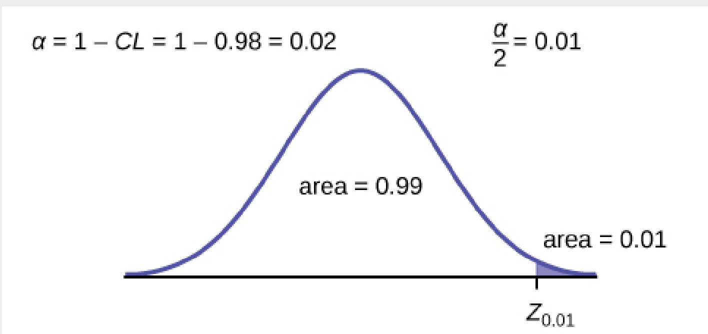
Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is σ = 0.337.

Solution:

Solution A
To find the confidence interval, start by finding the point estimate: the sample mean.

$$x = 1.024$$

Next, find the *EBM*. Because you are creating a 98% confidence interval, *CL* = 0.98.



You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01}$ = 2.326.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $x \pm EBM$.

x – *EBM* = 1.024 – 0.1431 = 0.8809

x – *EBM* = 1.024 – 0.1431 = 1.1671

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and

1.1671 watts per kilogram.

Solution:

Solution B

Note:

- Press STAT and arrow over to TESTS.
- Arrow down to 7:ZInterval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:
  - σ: 0.337
  - $\bar{x}$: 1.024
  - n: 30
  - C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

Note:
Try It
Exercise:

Problem:

[link] shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is σ = 0.337.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Solution:

$x = 0.940$

$\frac{\alpha}{2} = \frac{1-CL}{2} = \frac{1-0.93}{2} = 0.035$

$Z_{0.035} = 1.812$

$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$

$x - EBM = 0.940 - 0.1365 = 0.8035$

$x + EBM = 0.940 + 0.1365 = 1.0765$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in [link] and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example:
Exercise:

Problem:

Suppose we change the original problem in [link] by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution:

To find the confidence interval, you need the sample mean, x, and the *EBM*.

- x = 68
- *EBM* = $\left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; n = 36; The confidence level is 95% (CL = 0.95).

CL = 0.95 so α = 1 – CL = 1 – 0.95 = 0.05

$\frac{\alpha}{2} = 0.025 \quad z_{\frac{\alpha}{2}} = z_{0.025}$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$

x – EBM = 68 – 0.98 = 67.02

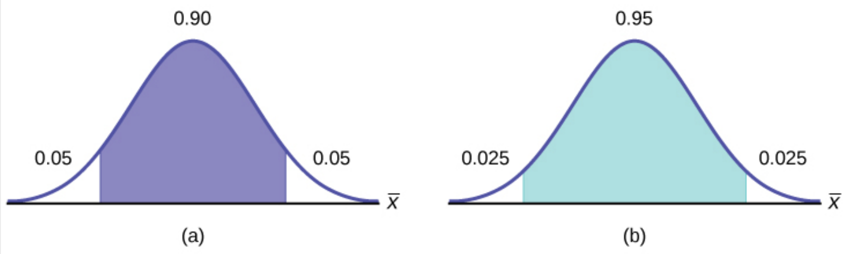
x + EBM = 68 + 0.98 = 68.98

Notice that the *EBM* is larger for a 95% confidence level in the original problem.

Interpretation
We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level
Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results
The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Note:
Try It
Exercise:

Problem:

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Solution:

(33.37, 38.63)

Example:
Suppose we change the original problem in [link] to see what happens to the error bound if the sample size is changed.
Exercise:

Problem:

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use *n* = 100 instead of *n* = 36? What happens if we decrease the sample size to *n* = 25 instead of *n* = 36?

- x = 68
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Solution:

Solution A
If we **increase** the sample size *n* to 100, we **decrease** the error bound.

When *n* = 100: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution:

Solution B
If we **decrease** the sample size *n* to 25, we **increase** the error bound.

When *n* = 25: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Note:
Try It
Exercise:

Problem:

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Solution:

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example:

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our

source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: *EBM* = 68.82 – 68 = 0.82.
- If we don't know the sample mean: *EBM* = $\frac{(68.82-67.18)}{2}$ = 0.82.

Calculate the Sample Mean:

- If we know the error bound: x = 68.82 – 0.82 = 68
- If we don't know the error bound: x = $\frac{(67.18+68.82)}{2}$ = 68.

Note:

Try It

Exercise:

Problem:

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Solution:

Sample mean is 45, error bound is 2.88

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is

$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$.

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n.

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example:
The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.

- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2} = 216.09$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Note:
Try It
Exercise:

Problem:

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Solution:

35 students

References

"American Fact Finder." U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).

"Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2, 2013).

"Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30,2013).

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/growthcharts/2000growthchart-us.pdf (accessed July 2, 2013).

La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).

"Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).

“Metadata Description of Candidate Summary File.” U.S. Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml (accessed July 2, 2013).

“National Health and Nutrition Examination Survey.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

Chapter Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for *n* to discover the size of the sample that is needed to achieve this goal:

$n = \frac{z^2\sigma^2}{EBM^2}$

Formula Review

$X \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by
(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)
= $(x - EBM, x + EBM)$
= $\left(x - z\frac{\sigma}{\sqrt{n}}, x + z\frac{\sigma}{\sqrt{n}}\right)$

EBM = $z\frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

α = 1 – *CL* = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the *z*-score with the property that the area to the right of the z-score is $\frac{\propto}{2}$ this is the *z*-score used in the calculation of *"EBM* where α = 1 – *CL*.

$n = \frac{z^2\sigma^2}{EBM^2}$ = the formula used to determine the sample size (*n*) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

(lower value, upper value) = (point estimate−error bound, point estimate + error bound)

To find the error bound when you know the confidence interval

error bound = upper value−point estimate OR error bound = $\frac{\text{upper value}-\text{lower value}}{2}$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format (x − EBM, x + EBM).

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

Exercise:

Problem: Identify the following:

a. x = ______
b. σ = ______
c. n = ______

Solution:

a. 244
b. 15
c. 50

Exercise:

Problem: In words, define the random variables X and X.

Exercise:

Problem: Which distribution should you use for this problem?

Solution:

$N\left(244, \frac{15}{\sqrt{50}}\right)$

Exercise:

Problem:

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

Exercise:

Problem:

What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Solution:

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

Exercise:

Problem: Identify the following:

a. x = ______
b. σ = ______
c. n = ______

Exercise:

Problem: In words, define the random variables X and X.

Solution:

X is the time in minutes it takes to complete the U.S. Census short form. X is the mean time it took a sample of 200 people to complete the U.S. Census short form.

Exercise:

Problem: Which distribution should you use for this problem?

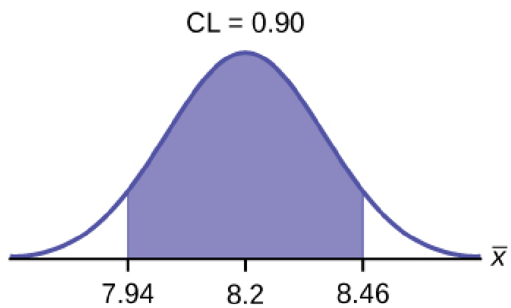
Exercise:

Problem:

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

Solution:

CI: (7.9441, 8.4559)



EBM = 0.26

Exercise:

Problem:

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

Exercise:

Problem:

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Solution:

The level of confidence would decrease because decreasing *n* makes the confidence interval wider, so at the same error bound, the confidence level decreases.

Exercise:

Problem:

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

Exercise:

Problem: Identify the following:

a. x = _______
b. σ = _______
c. n = _______

Solution:

a. $x = 2.2$
b. $\sigma = 0.2$
c. $n = 20$

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: In words, define the random variable $\bar{X}$.

Solution:

$\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

Exercise:

Problem: Which distribution should you use for this problem?

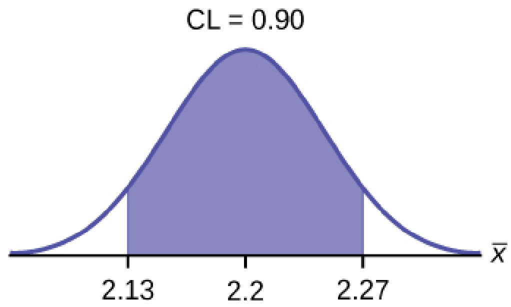
Exercise:

Problem:

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Solution:

EBM = 0.07
CI: (2.1264, 2.2736)



Exercise:

Problem:

Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Exercise:

Problem:

In complete sentences, explain why the confidence interval in [link] is larger than in [link].

Solution:

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

Exercise:

Problem:

In complete sentences, give an interpretation of what the interval in [link] means.

Exercise:

Problem:

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

Solution:

The confidence level would increase.

Exercise:

Problem:

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

Exercise:

Problem: x = _____

Solution:

30.4

Exercise:

Problem: n = _____

Exercise:

Problem: ________ = 15

Solution:

σ

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: What is x estimating?

Solution:

μ

Exercise:

Problem: Is σ_x known?

Exercise:

Problem:

As a result of your answer to [link], state the exact distribution to use when calculating the confidence interval.

Solution:

normal

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

Exercise:

Problem: How much area is in both tails (combined)? α =________

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2}$ =________

Solution:

0.025

Exercise:

Problem: Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

Exercise:

Problem: The 95% confidence interval is:____________________.

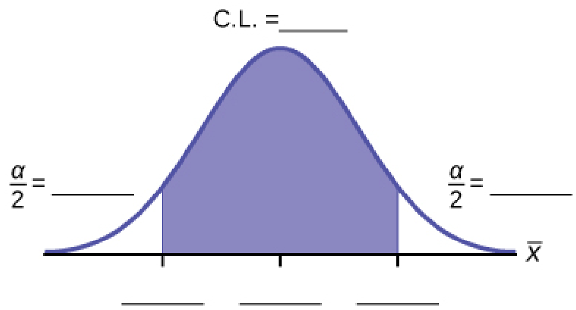
Solution:

(24.52,36.28)

Exercise:

Problem:

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Solution:

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

Exercise:

Problem:

Using the same mean, standard deviation, and level of confidence, suppose that *n* were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Solution:

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

Homework

Exercise:

Problem:

Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a.
   i. x =_________
   ii. σ =_________
   iii. n =_________

b. In words, define the random variables X and X.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean height of male Swedes.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

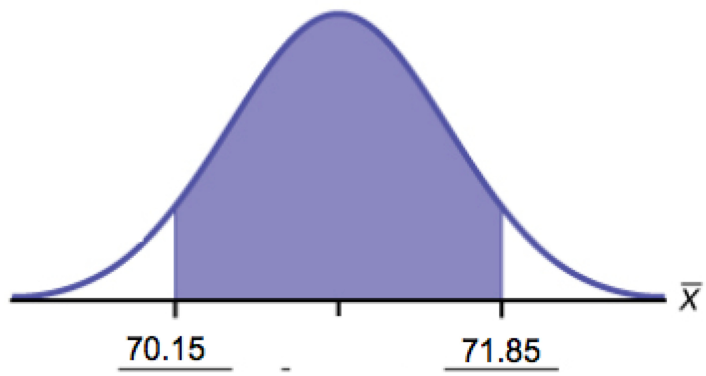
Solution:

a.
   i. 71
   ii. 3

iii. 48

b. X is the height of a Swiss male, and is the mean height from a sample of 48 Swiss males.
c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.49)

ii. 

iii. EBM = 0.849

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

Exercise:

Problem:

Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

a. In words, define the random variables X and X.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population mean length of engineering conferences.

i. State the confidence interval.

ii. Sketch the graph.
iii. Calculate the error bound.

Exercise:

Problem:

Suppose that an accounting firm does a study to determine the time needed to complete one person’s tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. i. x =_________
ii. σ =_________
iii. n =_________

b. In words, define the random variables X and X.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean time to complete the tax forms.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

Solution:

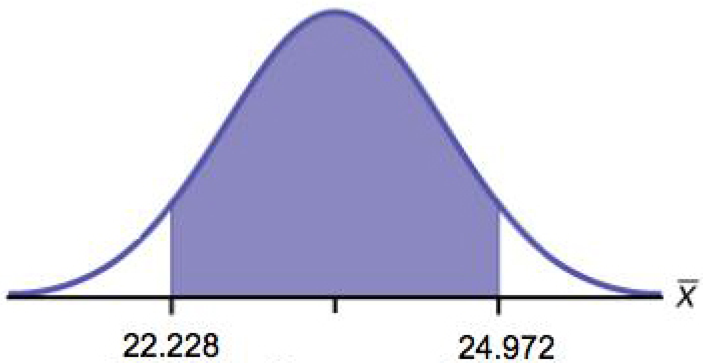
a. i. x = 23.6
ii. σ = 7

iii. n = 100

b. *X* is the time needed to complete an individual tax form. X is the mean time to complete tax forms from a sample of 100 customers.

c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.

d. i. (22.228, 24.972)



ii.

iii. *EBM* = 1.372

e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.

f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.

g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

Exercise:

Problem:

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. x =________

ii. σ =_________
iii. s_x =_________

b. In words, define the random variable X.
c. In words, define the random variable $\bar{X}$.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90% confidence interval for the population mean weight of the candies.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

f. Construct a 98% confidence interval for the population mean weight of the candies.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
h. In complete sentences, give an interpretation of what the interval in part f means.

Exercise:

Problem:

A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

a. i. $\bar{x}$ =_________
ii. σ =_________
iii. n =_________

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90% confidence interval for the population mean number of letters campers send home.

  i. State the confidence interval.
  ii. Sketch the graph.
  iii. Calculate the error bound.

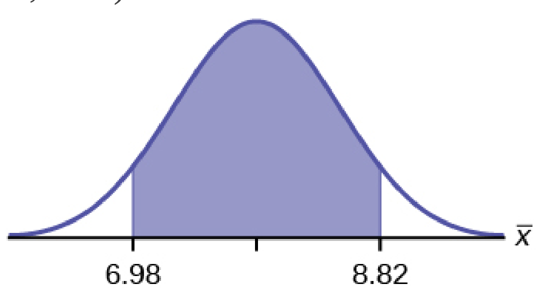
e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

Solution:

a.
  i. 7.9
  ii. 2.5
  iii. 20

b. X is the number of letters a single camper will send home. X is the mean number of letters sent home from a sample of 20 campers.

c. $N\ 7.9\left(\frac{2.5}{\sqrt{20}}\right)$

d.
  i. CI: (6.98, 8.82)
  ii. 

  iii. *EBM*: 0.92

e. The error bound and confidence interval will decrease.

Exercise:

Problem:

What is meant by the term “90% confident” when constructing a confidence interval for a mean?

a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

Exercise:

Problem:

The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. [link] shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200

$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the error bound.
c. Create a 95% confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

Solution:

a. x = $568,873
b. $CL = 0.95$ $\alpha = 1 – 0.95 = 0.05$ $z_{\frac{\alpha}{2}} = 1.96$
$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\ \frac{909200}{\sqrt{40}} = \$281{,}764$
c. $x - EBM = 568{,}873 - 281{,}764 = 287{,}109$
$x + EBM = 568{,}873 + 281{,}764 = 850{,}637$

Alternate solution:

> **Note:**
>
> 1. Press `STAT` and arrow over to `TESTS`.
> 2. Arrow down to `7:ZInterval`.
> 3. Press `ENTER`.
> 4. Arrow to Stats and press `ENTER`.
> 5. Arrow down and enter the following values:
>
>   - σ : 909,200
>   - x: 568,873

- n: 40
- CL: 0.95

6. Arrow down to Calculate and press `ENTER`.
7. The confidence interval is ($287,114, $850,632).
8. Notice the small difference between the two solutions–these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between $287,109 and $850,637.

Exercise:

Problem:

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

Exercise:

Problem:

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

Solution:

Use the formula for *EBM*, solved for *n*:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

From the statement of the problem, you know that $\sigma = 2.5$, and you need $EBM = 1$.

$z = z_{0.035} = 1.812$

(This is the value of z for which the area under the density curve to the ***right*** of z is 0.035.)

$$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$$

You need to measure at least 21 male students to achieve your goal.

Glossary

Confidence Level (CL)
: the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (*EBM*)
: the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

A Single Population Mean using the Student t Distribution

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation *s* as an estimate for *σ* and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing *σ* with *s* did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's t-distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever *s* is used as an estimate for *σ*.

If you draw a simple random sample of size *n* from a population that has an approximately normal distribution with mean *μ* and unknown population standard deviation *σ* and calculate the *t*-score $t = \frac{x-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the *t*-scores follow a **Student's t-distribution with *n* – 1 degrees of freedom**. The *t*-score has the same interpretation as the ***z*-score**. It measures how far x is from its mean *μ*. For each sample size *n*, there is a different Student's t-distribution.

The **degrees of freedom**, ***n* – 1**, come from the calculation of the sample standard deviation ***s***. In [link], we used *n* deviations $(x - x\text{values})$ to calculate ***s***. Because the sum of the deviations is zero, we can find the last deviation once we know the other ***n* – 1** deviations. The other ***n* – 1** deviations can change or vary freely. **We call the number *n* – 1 the degrees of freedom (df).**

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean *μ* and unknown population standard deviation *σ*. The size of the underlying population is generally not relevant unless it is very small. If

it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of *t*. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of *t* when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's t-Distribution.) When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t table (See [link]) gives *t*-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t-probabilities.**
The notation for the Student's t-distribution (using *T* as the random variable) is:

- $T \sim t_{df}$ where $df = n - 1$.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the **error bound for a population mean** is:

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is:
$(x - EBM, x + EBM)$.

Note:
To calculate the confidence interval directly:
Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example:
Exercise:

Problem:

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.
The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.
8.6 9.4 7.9 6.8 8.3 7.3 9.2 9.6 8.7 11.4 10.3 5.4 8.1 5.5 6.9

Solution:

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A
To find the confidence interval, you need the sample mean, x, and the *EBM*.

x = 8.2267 *s* = 1.6722 *n* = 15

df = 15 – 1 = 14 *CL* so *α* = 1 – *CL* = 1 – 0.95 = 0.05

$\frac{\alpha}{2}$ = 0.025 $t_{\frac{\alpha}{2}} = t_{0.025}$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

x – *EBM* = 8.2267 – 0.9240 = 7.3

x + *EBM* = 8.2267 + 0.9240 = 9.15

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution:

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).
Arrow to `Data` and press `ENTER`.
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq`.
Arrow down to `C-level` and enter 0.95
Arrow down to `Calculate` and press `ENTER`.
The 95% confidence interval is (7.3006, 9.1527)

Note:
Note
When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

Note:
Try It
Exercise:

Problem:

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Solution:

(8.1634, 9.8032)

Example:
Exercise:

Problem:

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. [link] shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution:

Solution A
From the sample, you can calculate x = 127.45 and s = 25.965. There are 20 infants in the sample, so n = 20, and df = 20 – 1 = 19.

You are asked to calculate a 90% confidence interval: CL = 0.90, so $\alpha = 1 - CL = 1 - 0.90 = 0.10$ $\frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05}$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is 1 – 0.05 = 0.95.

Use a table, calculator, or computer to find that $t_{0.05}$ = 1.729.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 1.729\left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$x - EBM$ = 127.45 – 10.038 = 117.412

$x + EBM$ = 127.45 + 10.038 = 137.488

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution:

Solution B

Note:
Enter the data as a list.
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to Data and press `ENTER`.
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90
Arrow down to `Calculate` and press `ENTER`.
The 90% confidence interval is (117.41, 137.49).

Note:
Try It
Exercise:

Problem:

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in [link]. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution:
Solution A

x= 6.133, *s* = 5.514, *n* = 15, and *df* = 15 – 1 = 14

CL = 0.98, so *α* = 1 - *CL* = 1 - 0.98 = 0.02

$\frac{\alpha}{2} = 0.01 t_{\frac{\alpha}{2}} = t_{0.01} = 2.624$

$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.624\left(\frac{5.514}{\sqrt{15}}\right)$~$3.736$

x – *EBM* = 6.133 – 3.736 = 2.397

x + *EBM* = 6.133 + 3.736 = 9.869

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution:
Solution B

Note:
Enter the data as a list.
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval`.
Press `ENTER`.
Arrow to `Data` and press `ENTER`.
Arrow down and enter the name of the list where the data is stored.
Enter `Freq`: 1
Enter `C-Level`: 0.98
Arrow down to `Calculate` and press `Enter`.
The 98% confidence interval is (2.3965, 9.8702).

References

"America's Best Small Companies." Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).

Data from *Microsoft Bookshelf*.

Data from http://www.businessweek.com/.

Data from http://www.forbes.com/.

“Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2,2013).

“Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn (accessed July 2, 2013).

“Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml (accessed July 2, 2013).

Chapter Review

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student’s t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{x - \mu}{{}^{s}/{\sqrt{n}}}$$

The t-score follows the Student’s t-distribution with n – 1 degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

Formula Review

s = the standard deviation of sample values.

$t = \frac{x-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student’s t-distribution

$df = n - 1$; the degrees of freedom for a Student’s t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student’s t-distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ = the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the *t*-score in the Student’s t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)
= (point estimate – *EBM*, point estimate + *EBM*)
= $\left(x - \frac{ts}{\sqrt{n}}, x + \frac{ts}{\sqrt{n}}\right)$

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

Exercise:

Problem: Identify the following:

a. x =________
b. s_x =________
c. n =________
d. n – 1 =________

Exercise:

Problem: Define the random variables *X* and X in words.

Solution:

X is the number of hours a patient waits in the emergency room before being called back to be examined. X is the mean wait time of 70 patients in the emergency room.

Exercise:

Problem: Which distribution should you use for this problem?

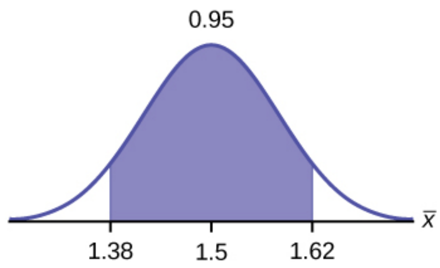
Exercise:

Problem:

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

Solution:

CI: (1.3808, 1.6192)



$EBM = 0.12$

Exercise:

Problem: Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

Exercise:

Problem: Identify the following:

a. x =________
b. s_x =________
c. n =________
d. $n – 1$ =________

Solution:

a. $x = 151$
b. $s_x = 32$
c. $n = 108$
d. $n – 1 = 107$

Exercise:

Problem: Define the random variable X in words.

Exercise:

Problem: Define the random variable X in words.

Solution:

X is the mean number of hours spent watching television per month from a sample of 108 Americans.

Exercise:

Problem: Which distribution should you use for this problem?

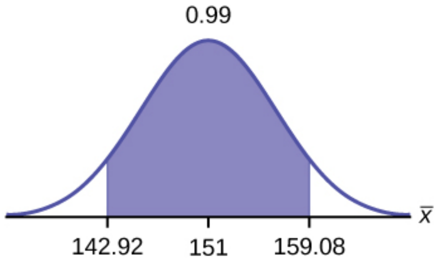
Exercise:

Problem:

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

Solution:

CI: (142.92, 159.08)



EBM = 8.08

Exercise:

Problem:

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in [link] are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Exercise:

Problem: Calculate the following:

a. x =_______
b. s_x =_______
c. n =_______

Solution:

a. 3.26
b. 1.02
c. 39

Exercise:

Problem: Define the random variable X in words.

Exercise:

Problem: What is x estimating?

Solution:

μ

Exercise:

Problem: Is σ_x known?

Exercise:

Problem:

As a result of your answer to [link], state the exact distribution to use when calculating the confidence interval.

Solution:

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

Exercise:

Problem: How much area is in both tails (combined)?

Exercise:

Problem: How much area is in each tail?

Solution:

0.025

Exercise:

Problem: Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise:

Problem: The 95% confidence interval is_____.

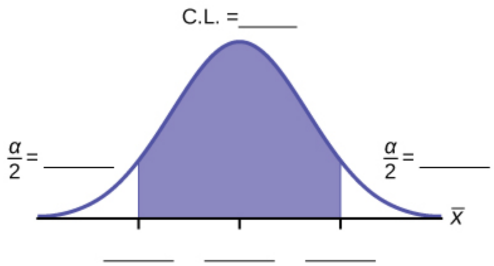
Solution:

(2.93, 3.59)

Exercise:

Problem:

Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Solution:

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

Exercise:

Problem:

Using the same x, s_x, and level of confidence, suppose that *n* were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Solution:

The error bound would become EBM = 0.245. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

Exercise:

Problem:

Using the same x, s_x, and *n* = 39, how would the error bound change if the confidence level were reduced to 90%? Why?

Homework

Exercise:

Problem:

In six packages of “The Flintstones® Real Fruit Snacks” there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

a. Define the random variables *X* and *P′* in words.

b. Which distribution should you use for this problem? Explain your choice
c. Calculate p'.
d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.

   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

Exercise:

Problem:

A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a. i. x = ___________
   ii. s_x = ___________
   iii. n = ___________
   iv. n – 1 = ___________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.

   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

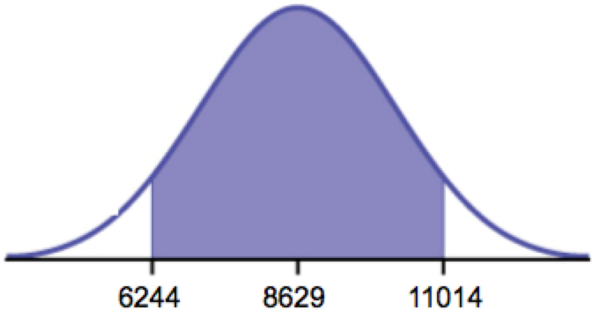
e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

Solution:

a. i. 8629
   ii. 6944
   iii. 35
   iv. 34

b. t_{34}

c. i. CI: (6244, 11,014)



ii.
iii. EB = 2385

d. It will become smaller

Exercise:

Problem:

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. i. x = ___________
ii. s_x = ___________
iii. n = ___________
iv. n – 1 = ___________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean time wasted.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

e. Explain in a complete sentence what the confidence interval means.

Exercise:

Problem:

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

a. i. x = ___________
ii. s_x = ___________
iii. n = ___________
iv. n – 1 = ___________

b. Define the random variable X in words.
c. Define the random variable X in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 95% confidence interval for the population mean length of time.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

f. What does it mean to be “95% confident” in this problem?

Solution:

a. i. x = 2.51
ii. s_x = 0.318
iii. n = 9
iv. n - 1 = 8

b. the effective length of time for a tranquilizer
c. the mean effective length of time of tranquilizers from a sample of nine patients
d. We need to use a Student’s-t distribution, because we do not know the population standard deviation.

e. i. CI: (2.27, 2.76)
ii. Check student's solution.
iii. *EBM*: 0.25

f. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

Exercise:

Problem:

Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. i. x = ___________
ii. s_x = ___________
iii. n = ___________

iv. $n - 1 =$ ___________

b. Define the random variable X in words.
c. Define the random variable X in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99% confidence interval for the population mean length of time using training wheels.

   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. Why would the error bound change if the confidence level were lowered to 90%?

Exercise:

Problem:

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80
$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

$x = \$251,854.23$

$s = \ \$521,130.41$

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

Solution:

$x = \$251,854.23$

$s = \ \$521,130.41$

Note that we are not given the population standard deviation, only the standard deviation of the sample.

There are 30 measures in the sample, so *n* = 30, and *df* = 30 - 1 = 29

CL = 0.96, so *α* = 1 - *CL* = 1 - 0.96 = 0.04

$\frac{\alpha}{2} = 0.02 t_{\frac{\alpha}{2}} = t_{0.02}$ = 2.150

$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.150 \left(\frac{521,130.41}{\sqrt{30}} \right)$ ~ $\$204,561.66$

x - *EBM* = \$251,854.23 - \$204,561.66 = \$47,292.57

x + *EBM* = \$251,854.23+ \$204,561.66 = \$456,415.89

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between \$47,292.57 and \$456,415.89.

Alternate Solution

Note:
Enter the data as a list.
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval`.
Press `ENTER`.
Arrow to Data and press `ENTER`.
Arrow down and enter the name of the list where the data is stored.
Enter `Freq`: 1
Enter `C-Level`: 0.96
Arrow down to `Calculate` and press `Enter`.
The 96% confidence interval is (\$47,262, \$456,447).

The difference between solutions arises from rounding differences.

Exercise:

Problem:

Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The [link] shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO’s for these top small firms. Use the Student's t-distribution.

Exercise:

Problem:

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

a.
  i. x = ____________
  ii. s_x = ____________
  iii. n = ____________
  iv. n-1 = ____________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.

    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

Solution:

a.
    i. x = 11.6
    ii. s_x = 4.1
    iii. *n* = 225
    iv. *n* - 1 = 224

b. *X* is the number of unoccupied seats on a single flight. X is the mean number of unoccupied seats from a sample of 225 flights.
c. We will use a Student's-t distribution, because we do not know the population standard deviation.

d.
    i. CI: (11.12 , 12.08)
    ii. Check student's solution.
    iii. *EBM*: 0.48

Exercise:

Problem:

In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable X in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.

    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

d. Explain what a "95% confidence interval" means for this study.

Exercise:

Problem:

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

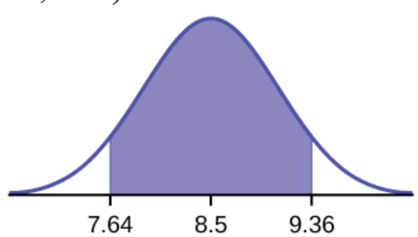
a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.

  i. State the confidence interval.
  ii. Sketch the graph.
  iii. Calculate the error bound.

b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

Solution:

a. i. CI: (7.64 , 9.36)

  ii. 

  iii. *EBM*: 0.86

b. The sample should have been increased.
c. Answers will vary.
d. Answers will vary.
e. Answers will vary.

Exercise:

Problem:

A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

a. i. x = ___________
  ii. s_x = ___________
  iii. n = ___________
  iv. n-1 = ___________

b. Define the random variables X and X in words.

c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean worth of coupons.

   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Exercise:

Problem:

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

Solution:

b

Exercise:

Problem: What is the error bound?

a. 0.87
b. 1.98
c. 0.99
d. 1.74

Glossary

Degrees of Freedom (*df*)

the number of objects in a sample that are free to vary

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu,\sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

Student's **t**-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n get larger.
- There is a "family of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

A Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03,0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying **distribution is a binomial distribution**. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success. To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\widehat{P}$, read "P hat".)

When n is large and p is not close to zero or one, we can use the **normal distribution** to approximate the binomial.

$$X \sim N(np, \sqrt{npq})$$

If we divide the random variable, the mean, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n.)

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify : $\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$

***P′* follows a normal distribution for proportions**: $\frac{X}{n} = P'\text{~}N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$

The confidence interval has the form (*p′* – *EBP*, *p′* + *EBP*). *EBP* is error bound for the proportion.

$p' = \frac{x}{n}$

p′ = the **estimated proportion** of successes (*p′* is a **point estimate** for *p*, the true proportion.)

x = the **number** of successes

n = the size of the sample

The error bound for a proportion is

$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$ where *q′* = 1 – *p′*

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{pq}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p'q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the **sample proportions *p′* and *q′* are estimates of the unknown population proportions *p* and *q***. The estimated proportions *p′* and *q′* are used because *p* and *q* are not known. The sample proportions *p′* and *q′* are calculated from the data: *p′* is the estimated proportion of successes, and *q′* is the estimated proportion of failures.

The confidence interval can be used only if the number of successes *np′* and the number of failures *nq′* are both greater than five.

Note:
Note
For the normal distribution of proportions, the *z*-score formula is as follows.

If $P'\text{~}N\left(p, \sqrt{\frac{pq}{n}}\right)$ then the *z*-score formula is $z = \frac{p'-p}{\sqrt{\frac{pq}{n}}}$

Example:
Exercise:

Problem:

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution:
Solution A

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Let *X* = the number of people in the sample who have cell phones. *X* is binomial. $X\text{~}B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find *p′*, *q′*, and *EBP*.

n = 500

x = the number of successes = 421

$p' = \frac{x}{n} = \frac{421}{500} = 0.842$

p′ = 0.842 is the sample proportion; this is the point estimate of the population proportion.

q′ = 1 – *p′* = 1 – 0.842 = 0.158

Since *CL* = 0.95, then *α* = 1 – *CL* = 1 – 0.95 = 0.05 $\left(\frac{\alpha}{2}\right)$ = 0.025.

Then $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p\prime – EBP = 0.842 – 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is (*p′* – *EBP*, *p′* + *EBP*) = (0.810, 0.874).

Interpretation
We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level
Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution:

Solution B

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to x and enter 421.
Arrow down to n and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.81003, 0.87397).

Note:

Try It
Exercise:

Problem:

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Solution:

(0.3315, 0.4525)

Example:
Exercise:

Problem:

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution:

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A
x = 300 and n = 500

$p' = \frac{x}{n} = \frac{300}{500} = 0.600$

$q' = 1 - p' = 1 - 0.600 = 0.400$

Since CL = 0.90, then α = 1 – CL = 1 – 0.90 = 0.10$\left(\frac{\alpha}{2}\right)$ = 0.05

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is

0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is (p' – EBP, p' + EBP) = (0.564,0.636).

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution:

Solution B

Note:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to x and enter 300.
Arrow down to n and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.564, 0.636).

Note:

Try It

Exercise:

Problem:

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

Solution:

(0.7731, 0.8269); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%.

Exercise:

Problem:

b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Solution:
Solution A

Sixty-eight percent (68%) of students own an iPod and a smart phone.

$p' = 0.68$

$q' = 1 – p' = 1 – 0.68 = 0.32$

Since CL = 0.97, we know α = 1 – 0.97 = 0.03 and $\frac{\alpha}{2}$ = 0.015.

The area to the left of $z_{0.015}$ is 0.015, and the area to the right of $z_{0.015}$ is 1 – 0.015 = 0.985.

Using the TI 83, 83+, or 84+ calculator function InvNorm(.985,0,1),

$z_{0.015} = 2.17$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = 2.17\sqrt{\frac{0.68(0.32)}{300}} \approx 0.0584$$

$p' - EPB = 0.68 - 0.0584 = 0.6216$

$p' + EPB = 0.68 + 0.0584 = 0.7384$

We are 97% confident that the true proportion of all students who own an iPod and a smart phone is between 0.6216 and 0.7384.

Solution:
Solution B

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 300*0.68.
Arrow down to n and enter 300.
Arrow down to C-Level and enter 0.97.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.6216, 0.7384).

"Plus Four" Confidence Interval for *p*

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example:

Exercise:

Problem:

A random sample of 25 statistics students was asked: “Have you smoked a cigarette in the past week?” Six students reported smoking within the past week. Use the “plus-four” method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution:
Solution A

Solution A
Six students out of 25 reported smoking within the past week, so x = 6 and n = 25. Because we are using the “plus-four” method, we will use x = 6 + 2 = 8 and n = 25 + 4 = 29.

$$p' = \frac{x}{n} = \frac{8}{29} \approx 0.276$$

$q' = 1 - p' = 1 - 0.276 = 0.724$

Since CL = 0.95, we know α = 1 – 0.95 = 0.05 and $\frac{\alpha}{2}$ = 0.025.

$z_{0.025} = 1.96$

$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{0.276(0.724)}{29}} \approx 0.163$

p' – EPB = 0.276 – 0.163 = 0.113

p' + EPB = 0.276 + 0.163 = 0.439

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Solution:

Solution B

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.

Note:
Reminder
Remember that the plus-four method assume an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

Arrow down to *x* and enter eight.
Arrow down to *n* and enter 29.
Arrow down to C-Level and enter 0.95.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.113, 0.439).

Note:
Try It
Exercise:

Problem:

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the "plus-four" method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution:
Solution A

Using "plus four," we have *x* = 31 + 2 = 33 and *n* = 65 + 4 = 69.

$p' = \frac{33}{69} \approx 0.478$

$q' = 1 - p' = 1 - 0.478 = 0.522$

Since *CL* = 0.96, we know *α* = 1 – 0.96 = 0.04 and $\frac{\alpha}{2}$ = 0.02.

$z_{0.02} = 2.054$

$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (2.054)\left(\sqrt{\frac{(0.478)(0.522)}{69}}\right)$~0.124

$p' – EPB = 0.478 – 0.124 = 0.354$

$p' + EPB = 0.478 + 0.124 = 0.602$

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

Solution:
Solution B

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to *x* and enter 33.
Arrow down to *n* and enter 69.
Arrow down to C-Level and enter 0.96.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.355, 0.602).

Example:
Exercise:

Problem:

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the "plus four" method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution:

Solution A
Using "plus-four," we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$.

$p' = \frac{15}{54} \approx 0.278$

$q' = 1 – p' = 1 - 0.241 = 0.722$

Since CL = 0.90, we know α = 1 – 0.90 = 0.10 and $\frac{\alpha}{2}$ = 0.05.

$z_{0.05} = 1.645$

$$EPB = (z_{\frac{\alpha}{2}})\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100$$

p' – EPB = 0.278 – 0.100 = 0.178

p' + EPB = 0.278 + 0.100 = 0.378

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook.

Solution:

Solution B

Note:

Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to *x* and enter 15.
Arrow down to *n* and enter 54.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.178, 0.378).

Note:
Try It
Exercise:

Problem:

The Berkman Center Study referenced in [link] talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the “plus-four” method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in [link].

Solution:
Solution A

Using “plus-four,” we have x = 159 + 2 = 161 and n = 588 + 4 = 592.

$p' = \frac{161}{592} \approx 0.272$

$q' = 1 - p' = 1 - 0.272 = 0.728$

Since CL = 0.90, we know α = 1 – 0.90 = 0.10 and $\frac{\alpha}{2}$ = 0.05

$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.272)(0.728)}{592}}\right) \approx 0.030$

p' – EPB = 0.272 – 0.030 = 0.242

p' + EPB = 0.272 + 0.030 = 0.302

We are 90% confident that between 24.2% and 30.2% of all teens would report having more than 500 friends on Facebook.

Solution:
Solution B

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 161.
Arrow down to n and enter 592.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.242, 0.302).

Conclusion: The confidence interval for the larger sample is narrower than the interval from [link]. Larger samples will always yield more precise confidence intervals than smaller samples. The "plus four" method has a greater impact on the smaller sample. It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the *EPB*, changing it from 0.102 to 0.100. In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592). It is easy to see that the plus-four method has the greatest impact on smaller samples.

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is

- $EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$
- Solving for *n* gives you an equation for the sample size.
- $n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2}$

Example:
Exercise:

Problem:

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Solution:

From the problem, we know that ***EBP* = 0.03** (3%=0.03) and $z_{\frac{\alpha}{2}}$ $z_{0.05}$ = 1.645 because the confidence level is 90%.

However, in order to find *n*, we need to know the estimated (sample) proportion *p*′. Remember that *q*′ = 1 – *p*′. But, we do not know *p*′ yet. Since we multiply *p*′ and *q*′

together, we make them both equal to 0.5 because $p'q'$ = (0.5)(0.5) = 0.25 results in the largest possible product. (Try other products: (0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16 and so on). The largest possible product gives us the largest *n*. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size *n*, use the formula and make the substitutions.

$$n = \frac{z^2 p'q'}{EBP^2} \text{ gives } n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Note:
Try It
Exercise:

Problem:

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Solution:

271 customers should be surveyed.

References

Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons.” Public Policy Polling. Available online at http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. “Teens, Social Media, and Privacy.” PewInternet, 2013. Available online at http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx (accessed July 2, 2013).

Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Saad, Lydia. “Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity.” Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx (accessed July 2, 2013).

The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).

Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll (accessed July 2, 2013).

“52% Say Big-Time College Athletics Corrupt Education Process.” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

Chapter Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}}\right)$$

The “plus four” method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n+4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated

to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = x / n$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$q' = 1 – p'$

$p' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ The variable p' has a binomial distribution that can be approximated with the normal distribution shown here.

EBP = the error bound for a proportion = $z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}$

Confidence interval for a proportion:

$$(\text{lower bound, upper bound}) = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}},\ p' + z\sqrt{\frac{p'q'}{n}}\right)$$

$n = \frac{z_{\frac{\alpha}{2}}{}^2 p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence 1 - α and margin of error EBP.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}}\ p'+q'=1$$

The confidence interval has the format (p' – EBP, p' + EBP).

x is a point estimate for μ

p' is a point estimate for ρ

s is a point estimate for σ

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

Exercise:

Problem:

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

Exercise:

Problem:

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Solution:

It would decrease, because the z-score would decrease, which reducing the numerator and lowering the number.

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

Exercise:

Problem: Identify the following:

a. x = _______
b. n = _______
c. p' = _______

Exercise:

Problem: Define the random variables X and P' in words.

Solution:

X is the number of “successes” where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

Exercise:

Problem: Which distribution should you use for this problem?

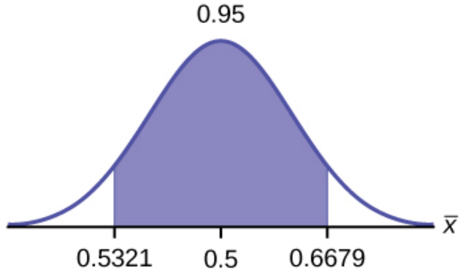
Exercise:

Problem:

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

Solution:

CI: (0.5321, 0.6679)



EBM: 0.0679

Exercise:

Problem:

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

Exercise:

Problem:

We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables *X* and *P′* in words.

Solution:

X is the number of “successes” where an executive prefers a truck. *P′* is the percentage of executives sampled who prefer a truck.

Exercise:

Problem: Which distribution should you use for this problem?

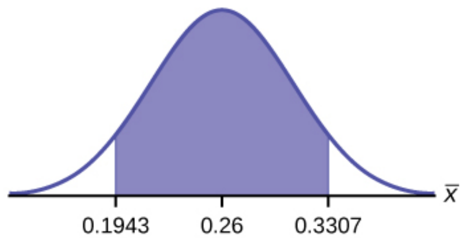
Exercise:

Problem:

Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

Solution:

CI: (0.19432, 0.33068)



EBM: 0.0707

Exercise:

Problem:

Suppose we want to lower the sampling error. What is one way to accomplish that?

Exercise:

Problem:

The sampling error given in the survey is ±2%. Explain what the ±2% means.

Solution:

The sampling error means that the true mean can be 2% above or below the sample mean.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent

answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

Exercise:

Problem: Define the random variable X in words.

Exercise:

Problem: Define the random variable P' in words.

Solution:

P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 90% confidence interval, and state the confidence interval and the error bound.

Solution:

CI: (0.62735, 0.67265)

EBM: 0.02265

Exercise:

Problem:

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

Exercise:

Problem: What is being counted?

Solution:

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: Calculate the following:

a. x = ________
b. n = ________
c. p' = ________

Solution:

a. $x = 64$
b. $n = 80$
c. $p' = 0.8$

Exercise:

Problem: State the estimated distribution of X. X~________

Exercise:

Problem: Define a new random variable P'. What is p' estimating?

Solution:

p

Exercise:

Problem: In words, define the random variable P'.

Exercise:

Problem:

State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

Solution:

$P' \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$. (0.72171, 0.87829).

Exercise:

Problem: How much area is in both tails (combined)?

Exercise:

Problem: How much area is in each tail?

Solution:

0.04

Exercise:

Problem: Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise:

Problem: The 92% confidence interval is _______.

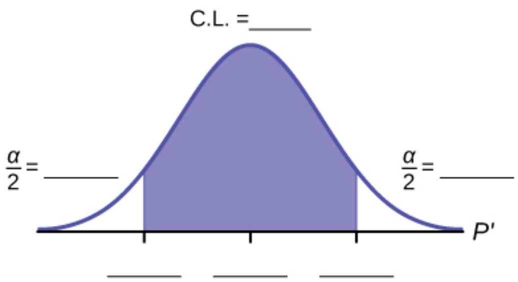
Solution:

(0.72; 0.88)

Exercise:

Problem:

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Solution:

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

Exercise:

Problem:

Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Solution:

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

Exercise:

Problem:

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Homework

Exercise:

Problem:

Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Solution:

a. 1,068
b. The sample size would need to be increased since the critical value increases as the confidence level increases.

Exercise:

Problem:

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a.
   i. x = ____________
   ii. n = ____________
   iii. p' = ____________

b. Define the random variables X and P', in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

Exercise:

Problem:

According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.

    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

Solution:

a. X = the number of people who feel that the president is doing an acceptable job;

    P' = the proportion of people in a sample who feel that the president is doing an acceptable job.
b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
c.
    i. CI: (0.59, 0.63)
    ii. Check student's solution
    iii. EBM: 0.02

Exercise:

Problem:

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

a. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families. Define the random variables X and P', in words.
b. Which distribution should you use for this problem? Explain your choice.

c. Construct a 95% confidence interval.

    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

Exercise:

Problem: Refer to the information in [link].

a. Construct three 95% confidence intervals.

    i. percent of all Asians who would welcome a white person into their families.
    ii. percent of all Asians who would welcome a Latino into their families.
    iii. percent of all Asians who would welcome a black person into their families.

b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

Solution:

a.
    i. (0.72, 0.82)
    ii. (0.65, 0.76)
    iii. (0.60, 0.72)

b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

Exercise:

Problem:

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.

    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

d. Explain what a “97% confidence interval” means for this study.

Exercise:

Problem:

A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.

    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

Solution:

a. X = the number of adult Americans who feel that crime is the main problem; P' = the proportion of adult Americans who feel that crime is the main problem
b. Since we are estimating a proportion, given P' = 0.2 and n = 1000, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c. i. CI: (0.18, 0.22)
ii. Check student's solution.
iii. *EBM*: 0.02

d. One way to lower the sampling error is to increase the sample size.
e. The stated "± 3%" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

Exercise:

Problem:

Refer to [link]. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to

construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

Exercise:

Problem: A point estimate for the true population proportion is:

a. 0.90
b. 1.27
c. 0.79
d. 400

Solution:

c

Exercise:

Problem: A 90% confidence interval for the population proportion is _______.

a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

Exercise:

Problem: The error bound is approximately _____.

a. 1.581
b. 0.791
c. 0.059
d. 0.030

Solution:

d

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred

seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

Exercise:

Problem:

Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

Exercise:

Problem:

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is ______.

a. 0.6614
b. 0.3386
c. 173
d. 338

Solution:

a

Exercise:

Problem:

On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ±3% margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify *CL* and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.

g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

Exercise:

Problem:

A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from part a and n = 1,000 to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

Solution:

a. p' = $\frac{(0.55+0.49)}{2}$ = 0.52; EBP = 0.55 - 0.52 = 0.03
b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
c. CL = 0.75, so α = 1 – 0.75 = 0.25 and $\frac{\alpha}{2} = 0.125$ $z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is 1 – 0.125 = 0.875.)
   $EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1{,}000}} \approx 0.018$
   (p' - EBP, p' + EBP) = (0.52 – 0.018, 0.52 + 0.018) = (0.502, 0.538)

   Alternate Solution

   Note:
   STAT TESTS A: 1-PropZinterval with x = (0.52)(1,000), n = 1,000, CL = 0.75. Answer is (0.502, 0.538)

d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

Exercise:

Problem:

Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

Exercise:

Problem:

You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

Solution:

$CL = 0.95$ $\alpha = 1 - 0.95 = 0.05$ $\frac{\alpha}{2} = 0.025$ $z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$$n = \frac{{z_{\frac{\alpha}{2}}}^2 p'q'}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16$$

You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

Exercise:

Problem:

In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as "likely" or "very likely." Use the "plus four" method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

Glossary

Binomial Distribution

a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(\mathbf{n},\mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Error Bound for a Population Proportion (EBP)

the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Introduction
class="introduction"

You can use a hypothesis test to decide if a dog breeder’s claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "**hypothesis testing**." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note:
Note
To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

Glossary

Confidence Interval (CI)
: an interval estimate for an unknown population parameter. This depends on:

  - The desired confidence level.
  - Information that is known about the distribution (for example, known standard deviation).
  - The sample and its size.

Hypothesis Testing
: Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis:** It is a statement of no difference between sample means or proportions or no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)

H_0	H_a
less than or equal to (≤)	more than (>)

Note:
Note
H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

Example:
H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Note:
Try It
Exercise:

Problem:

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Solution:

H_0 : The drug reduces cholesterol by 25%. $p = 0.25$

H_a : The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Example:
We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:
H_0: $\mu = 2.0$
H_a: $\mu \neq 2.0$

Note:
Try It
Exercise:

Problem:

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ __ 66
b. H_a: μ __ 66

Solution:

a. H_0 : $\mu = 66$
b. H_a : $\mu \neq 66$

Example:

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:
H_0: $\mu \geq 5$
H_a: $\mu < 5$

Note:
Try It
Exercise:

Problem:

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ __ 45
b. H_a: μ __ 45

Solution:

a. H_0: $\mu \geq 45$
b. H_a: $\mu < 45$

Example:
In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.
H_0: $p \leq 0.066$
H_a: $p > 0.066$

Note:
Try It
Exercise:

Problem:

On a state driver’s test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: p __ 0.40
b. H_a: p __ 0.40

Solution:

a. H_0: $p = 0.40$
b. H_a: $p > 0.40$

Note:
Collaborative Exercise
Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Chapter Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤ or ≥)
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., (≠, >, or <).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_o has:	equal (=)	greater than or equal to (≥)	less than or equal to (≤)
then H_a has:	not equal (≠) **or** greater than (>) **or** less than (<)	less than (<)	greater than (>)

If $\alpha \le p$-value, then do not reject H_0.

If $\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

Exercise:

Problem:

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

Solution:

The random variable is the mean Internet speed in Megabits per second.

Exercise:

Problem:

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

Exercise:

Problem:

The American family has an average of two children. What is the random variable? Describe in words.

Solution:

The random variable is the mean number of children an American family has.

Exercise:

Problem:

The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

Exercise:

Problem:

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

Solution:

The random variable is the proportion of people picked at random in Times Square visiting the city.

Exercise:

Problem:

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

Exercise:

Problem:

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

Solution:

a. H_0: $p = 0.42$
b. H_a: $p < 0.42$

Exercise:

Problem:

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0: _________
b. H_a: _________

Exercise:

Problem:

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0: ___________
b. H_a: ___________

Solution:

a. H_0: $\mu = 15$
b. H_a: $\mu \neq 15$

Exercise:

Problem:

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: _________
b. H_a: _________

Homework

Exercise:

Problem:

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.

g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

Solution:

a. H_0: $\mu = 34$; H_a: $\mu \neq 34$
b. H_0: $p \leq 0.60$; H_a: $p > 0.60$
c. H_0: $\mu \geq 100{,}000$; H_a: $\mu < 100{,}000$
d. H_0: $p = 0.29$; H_a: $p \neq 0.29$
e. H_0: $p = 0.05$; H_a: $p < 0.05$
f. H_0: $\mu \leq 10$; H_a: $\mu > 10$
g. H_0: $p = 0.50$; H_a: $p \neq 0.50$
h. H_0: $\mu = 6$; H_a: $\mu \neq 6$
i. H_0: $p \geq 0.11$; H_a: $p < 0.11$
j. H_0: $\mu \leq 20{,}000$; H_a: $\mu > 20{,}000$

Exercise:

Problem:

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

Exercise:

Problem:

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

Solution:

c

Exercise:

Problem:

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. H_o: = 4.5, H_a : > 4.5
b. H_o: $\mu \geq 4.5$, H_a: $\mu < 4.5$
c. H_o: $\mu = 4.75$, H_a: $\mu > 4.75$
d. H_o: $\mu = 4.5$, H_a: $\mu > 4.5$

References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis
: a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	**H_0 IS ACTUALLY**	**...**
	True	False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject H_0** when **H_0 is true (correct decision).**
2. The decision is to **reject H_0** when **H_0 is true** (incorrect decision known as a**Type I error**).
3. The decision is **not to reject H_0** when, in fact, **H_0 is false** (incorrect decision known as a **Type II error**).
4. The decision is to **reject H_0** when **H_0 is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = ***P*(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = ***P*(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The Power of the Test is $1 – \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example:
Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.
Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.
α = probability that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. **β = probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.
Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Note:
Try It
Exercise:

Problem:

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Solution:

Type I error: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.

Type II error: The researcher thinks the blood cultures do not contain traces of pathogen *X*, when in fact, they do.

Example:

Suppose the null hypothesis, H_0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

***α* = probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error). ***β* = probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Note:

Try It

Exercise:

Problem:

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Solution:

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Example:
It’s a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It’s a Boy Genetic Labs has no effect on gender outcome.
Type I error: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It’s a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.
Type II error: This results when we fail to reject a false null hypothesis. In context, we would state that It’s a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.
The error of greater consequence would be the Type I error since couples would use the It’s a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Note:
Try It
Exercise:

Problem:

“Red tide” is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Solution:

In this scenario, an appropriate null hypothesis would beH_0: the mean level of toxins is at most 800 μg, $H_0 : \mu_0 \leq 800$ μg.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800 μg. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than 800 μg). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example:
A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?
Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
Type II: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.
In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient’s (and doctor’s) choice about whether to use the drug as a treatment option.

Note:

Try It

Determine both Type I and Type II errors for the following scenario: Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%.

Exercise:

Problem:

Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Solution:

Type I error: c

Type I error: b

Chapter Review

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

Exercise:

Problem:

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Solution:

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

Exercise:

Problem:

A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

Exercise:

Problem: For Exercise 9.12, what are α and β in words?

Solution:

α = the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

β = the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

Exercise:

Problem: In words, describe 1 – β For Exercise 9.12.

Exercise:

Problem:

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

Solution:

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

Exercise:

Problem:

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

Exercise:

Problem:

The power of a test is 0.981. What is the probability of a Type II error?

Solution:

0.019

Exercise:

Problem:

A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

Exercise:

Problem:

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

Solution:

0.998

Exercise:

Problem:

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

Homework

Exercise:

Problem:

State the Type I and Type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

Solution:

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults

ride the bus to work in Los Angeles when, in fact, fewer that 5% do.

f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

Exercise:

Problem:

For statements a-j in Exercise 9.109, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

Exercise:

Problem:

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

a. To conclude the drug is safe when in, fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

Solution:

b

Exercise:

Problem:

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is ________.

a. at least 20%, when in fact, it is less than 20%.
b. 20%, when in fact, it is 20%.
c. less than 20%, when in fact, it is at least 20%.
d. less than 20%, when in fact, it is less than 20%.

Exercise:

Problem:

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

Solution:

d

Exercise:

Problem:

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same

c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

Glossary

Type 1 Error
: The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error
: The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a **normal distribution** or a **Student's *t*-distribution**. (Remember, use a Student's *t*-distribution when the population **standard deviation** is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually *n* is large).

If you are testing a **single population mean**, the distribution for the test is for **means**:

$$X \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \text{ or } t_{df}$$

The population parameter is μ. The estimated value (point estimate) for μ is x, the sample mean.

If you are testing a **single population proportion**, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

The population parameter is *p*. The estimated value (point estimate) for *p* is *p′*. $p' = \frac{x}{n}$ where *x* is the number of successes and *n* is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean *μ*** using a **Student's *t*-distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean μ** using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion p**, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Chapter Review

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and

$nq > n$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α, then use $\alpha = 0.05$.
Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's *t*-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_x$ and $\sigma_x = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = \boldsymbol{p}$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Exercise:

Problem:

Which two distributions can you use for hypothesis testing for this chapter?

Solution:

A normal distribution or a Student’s *t*-distribution

Exercise:

Problem:

Which distribution do you use when you are testing a population mean and the population standard deviation is known? Assume a normal distribution, with $n \geq 30$.

Exercise:

Problem:

Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

Solution:

Use a Student's *t*-distribution

Exercise:

Problem:

A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

Exercise:

Problem:

A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

Solution:

a normal distribution for a single population mean

Exercise:

Problem:

It is thought that 42% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39% preferred Brand *A*. What distribution should you use to perform a hypothesis test?

Exercise:

Problem:

You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. What must you assume about the distribution of the data?

Solution:

It must be approximately normally distributed.

Exercise:

Problem:

You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

Exercise:

Problem:

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of *np* and *nq*?

Solution:

They must both be greater than five.

Exercise:

Problem:

You are performing a hypothesis test of a single population proportion. You find out that *np* is less than five. What must you do to be able to perform a valid hypothesis test?

Exercise:

Problem:

You are performing a hypothesis test of a single population proportion. The data come from which distribution?

Solution:

binomial distribution

Homework

Exercise:

Problem:

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is X ~ ________________

a. $N(7.24, \frac{1.93}{\sqrt{22}})$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

Solution:

d

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, *n*, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in *n* trials. The notation is: $X \sim B(n, p)$ $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Student's *t*-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption**—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

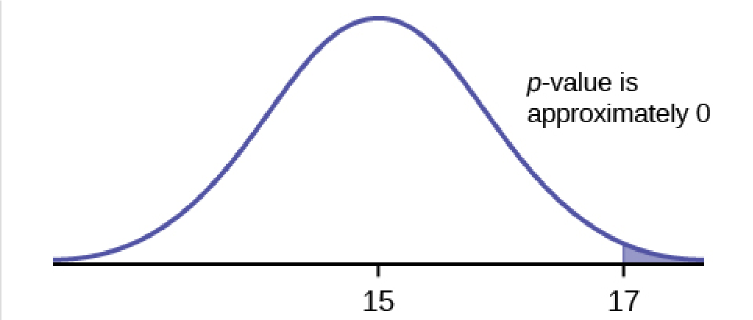
Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the ***p*-value**. The *p*-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large *p*-value calculated from the data indicates that we should not reject the **null hypothesis**. The smaller the *p*-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the *p*-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example:
Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm. and the distribution of heights is normal.
The null hypothesis could be H_0: $\mu \leq 15$ The alternate hypothesis is H_a: $\mu > 15$
The words **"is more than"** translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.
Since **σ is known** ($\sigma = 0.5$ cm.), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.
Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The *p*-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.
The *p*-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.



p-value= $P(x > 17)$ which is approximately zero.
A *p*-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Note:
Try It
Exercise:

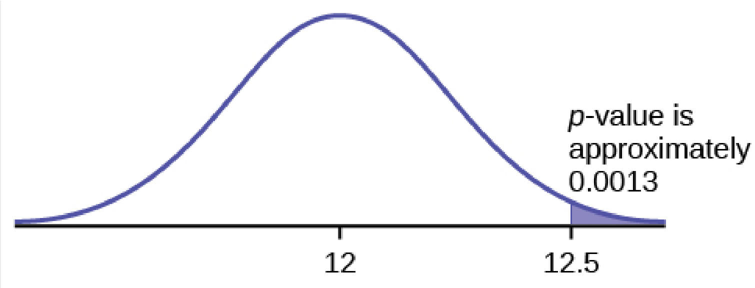
Problem:

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

H_0: $\mu \leq 12$
H_a: $\mu > 12$
The *p*-value is 0.0013
Draw a graph that shows the *p*-value.

Solution:

p-value = 0.0013



Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the *p*-value and a **preset or preconceived α (also called a "significance level")**. A preset *α* is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If $\alpha \leq p$-value, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis,H_a, may be correct.
- When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_o.

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Example:

When using the *p*-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device
If the *p*-value is low, the null must go.
If the *p*-value is high, the null must fly.
This memory aid relates a *p*-value less than the established alpha (the *p* is low) as rejecting the null hypothesis and, likewise, relates a *p*-value higher than the established alpha (the *p* is high) as not rejecting the null hypothesis.

Exercise:

Problem: Fill in the blanks.

Reject the null hypothesis when ____.

The results of the sample data _______________________________________.

Do not reject the null when hypothesis when ____.

The results of the sample data ____.

Solution:

Reject the null hypothesis when **the *p*-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the *p*-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Note:

Try It
Exercise:

Problem:

It’s a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

H_0: p = 0.50, H_a: p > 0.50

α = 0.01

p-value = 0.025

Interpret the results and state a conclusion in simple, non-technical terms.

Solution:

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support It’s a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

Chapter Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

1. $\alpha > p$-value, reject the null hypothesis
2. $\alpha \leq p$-value, do not reject the null hypothesis

Exercise:

Problem: When do you reject the null hypothesis?

Exercise:

Problem:

The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

Solution:

The outcome of winning is very unlikely.

Exercise:

Problem:

The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

Exercise:

Problem:

It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The *p*-value is almost zero. State the null and alternative hypotheses and interpret the *p*-value.

Solution:

H_0: $\mu > = 73$
H_a: $\mu < 73$
The *p*-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play

basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

Exercise:

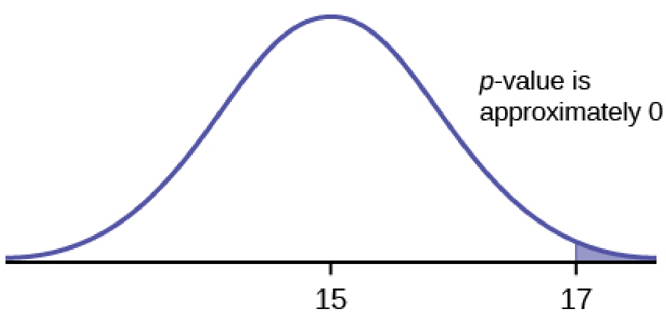
Problem:

The mean age of graduate students at a University is at most 31 y ears with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The *p*-value is 0.0264. State the null and alternative hypotheses and interpret the *p*-value.

Exercise:

Problem:

Does the shaded region represent a low or a high *p*-value compared to a level of significance of 1%?



Solution:

The shaded region shows a low *p*-value.

Exercise:

Problem: What should you do when α > *p*-value?

Exercise:

Problem: What should you do if α = *p*-value?

Solution:

Do not reject H_0.

Exercise:

Problem:

If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Exercise:

Problem: Is this a test of means or proportions?

Solution:

means

Exercise:

Problem: What symbol represents the random variable for this test?

Exercise:

Problem: In words, define the random variable for this test.

Solution:

the mean time spent in jail for 26 first time convicted burglars

Exercise:

Problem: Is σ known and, if so, what is it?

Exercise:

Problem: Calculate the following:

a. x ________
b. σ ________
c. s_x ________
d. n ________

Solution:

a. 3
b. 1.5
c. 1.8
d. 26

Exercise:

Problem:

Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

Exercise:

Problem: State the distribution to use for the hypothesis test.

Solution:

$X \sim N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

Exercise:

Problem:

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: ______________________ H_a : ______________________
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Calculate the following:

   i. x = ______________
   ii. s = ______________
   iii. n = ______________

h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:

   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

Homework

Exercise:

Problem:

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: ______________________ H_a: ______________________
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:

   i. x = __________________
   ii. n = __________________
   iii. p' = _______________

g. Calculate σ_x = ___________. Show the formula set-up.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived $\alpha = 0.05$, what is your:

   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

Glossary

Level of Significance of the Test

probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

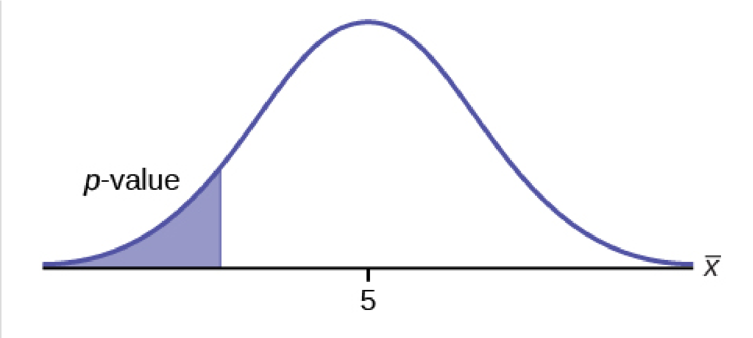
p-value
: the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the *p*-value, the stronger the evidence is against the null hypothesis.

Additional Information and Full Hypothesis Test Examples

- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, a common standard to use is α = 0.05.**
- When you calculate the *p*-value and draw the picture, the *p*-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternative hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the *p*-value**: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller *p*-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large *p*-value such as 0.4, as opposed to a *p*-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example:
H_o: μ = 5, H_a: μ < 5
Test of a single population mean. H_a tells you the test is left-tailed. The picture of the *p*-value is as follows:



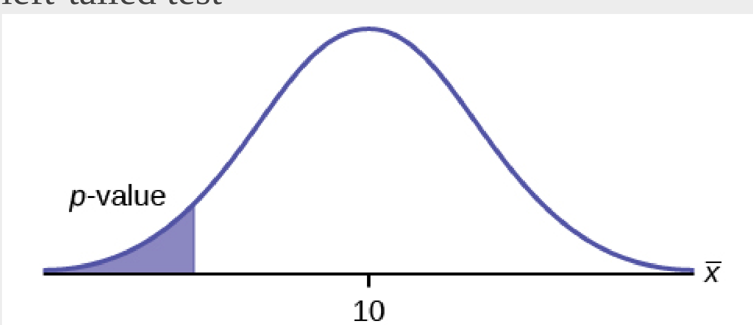
Note:
Try It
Exercise:

Problem: H_0: $\mu = 10$, H_a: $\mu < 10$

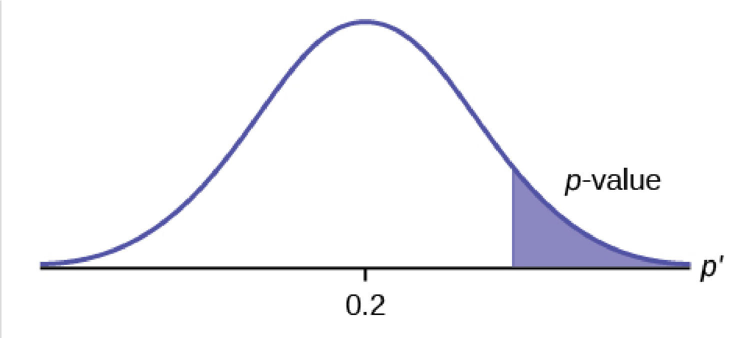
Assume the *p*-value is 0.0935. What type of test is this? Draw the picture of the *p*-value.

Solution:

left-tailed test



Example:
H_0: $p \leq 0.2$ H_a: $p > 0.2$
This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the *p*-value is as follows:



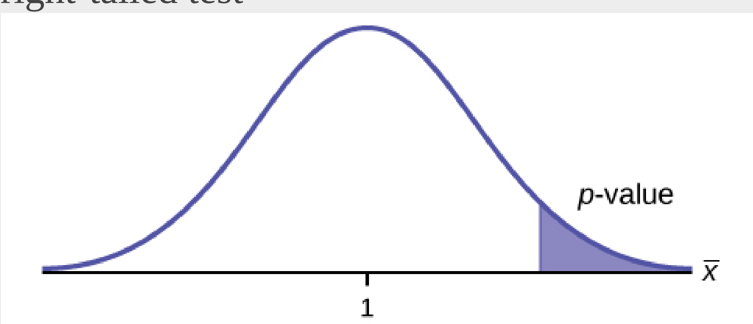
Note:
Try It
Exercise:

Problem: H_0: $\mu \leq 1$, H_a: $\mu > 1$

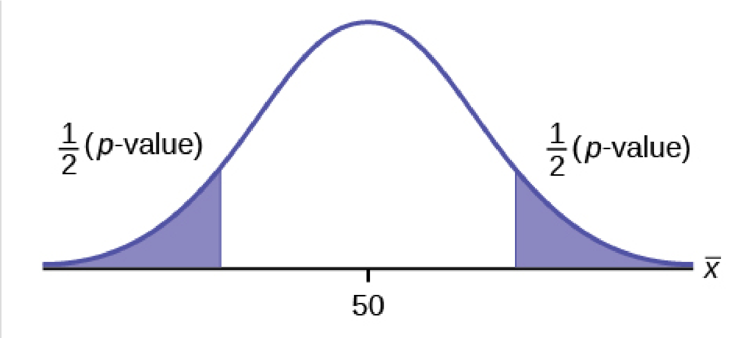
Assume the *p*-value is 0.1243. What type of test is this? Draw the picture of the *p*-value.

Solution:

right-tailed test



Example:
H_0: $p = 50$ H_a: $p \neq 50$
This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the *p*-value is as follows.



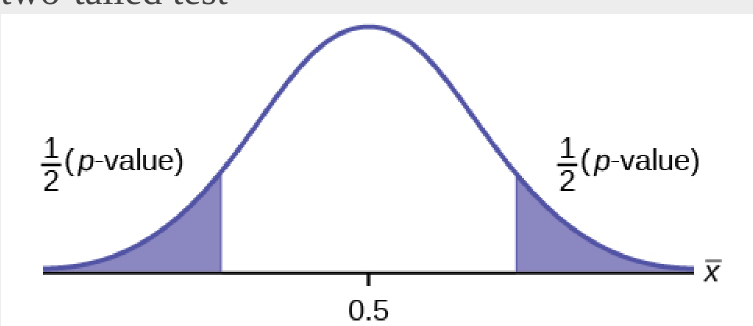
Note:
Try It
Exercise:

Problem: H_0: $p = 0.5$, H_a: $p \neq 0.5$

Assume the p-value is 0.2564. What type of test is this? Draw the picture of the p-value.

Solution:

two-tailed test



Full Hypothesis Test Examples

Example:
Exercise:

Problem:

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

H_0: $\mu = 16.43$ H_a: $\mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population **standard deviation** is known: $\sigma = 0.8$)

$\bar{X} \sim N\left(\mu, \frac{\sigma_X}{\sqrt{n}}\right)$ Therefore, $\bar{X} \sim N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

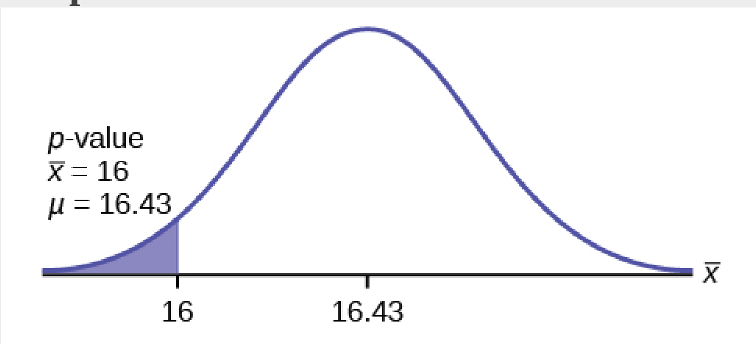
$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the *p*-value using the normal distribution for a mean:

p-value = $P(x < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value = 0.0187 (This is called the **actual level of significance**.) The *p*-value is the area to the left of the sample mean is given as 16.

Graph:



$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the *p*-value: If H_0 is true, there is a 0.0187 probability (1.87%)that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the *p*-value:

$\alpha = 0.05$ *p*-value = 0.0187 $\alpha >$ *p*-value

Make a decision: Since $\alpha >$ *p*-value, reject H_0.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is

sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The *p*-value can easily be calculated.

Note:
Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for *n*. Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the *p*-value (p = 0.0187) but it also calculates the test statistic (*z*-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw`(instead of `Calculate`). Press `ENTER`. A shaded graph appears with z = -2.08 (test statistic) and p = 0.0187 (*p*-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

When the calculator does a *Z*-Test, the `Z-Test` function finds the *p*-value by doing a normal probability calculation using the **central limit theorem**:

$P(x < 16) =$ `2nd DISTR normcdf`
$\left(-10 \text{ ^ } 99, 16, 16.43, 0.8/\sqrt{15}\right)$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Note:
Try It
Exercise:

Problem:

The mean throwing distance of a football for Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Note:
Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for $\mu 0$ (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as <, ≠, or >. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value but it also calculates the test statistic (z-score) for the sample mean. Select <, ≠, or > for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and p-value. Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

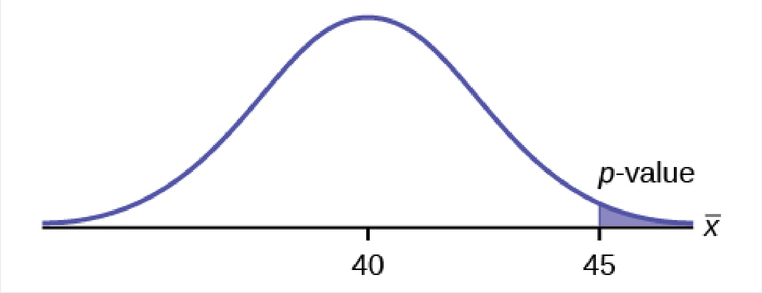
Solution:

Since the problem is about a mean, this is a test of a single population mean.

$H_0 : \mu = 40$

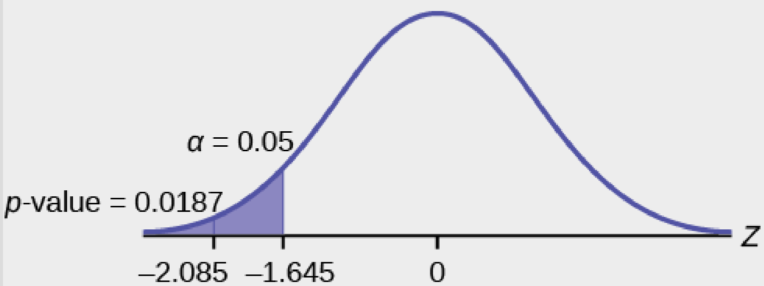
$H_a : \mu > 40$

$p = 2.6115 \times 10^{-29}$



Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Note:
Historical Note ([link])
The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{x - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$. For this problem, x = 16, μ_X = 16.43 from the null hypothes is, σ_X = 0.8, and n = 15.) You can find the critical value for α = 0.05 in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that α > p-value), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187 is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the p-value, and the test statistic and the critical value.



Example:
Exercise:

Problem:

A college football coach records the mean weight that his players can bench press as **275 pounds**, with a **standard deviation of 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3) 215(3) 225(1) 241(2) 252(2) 265(2) 275(2) 313(2) 316(5) 338(2) 341(1) 345(2) 368(2) 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is **more than 275 pounds**.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a **test of a single population mean**.

H_0: $\mu = 275$
H_a: $\mu > 275$
This is a right-tailed test.

Calculating the distribution needed:

Random variable: X = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$X \sim N\left(275, \frac{55}{\sqrt{30}}\right)$$

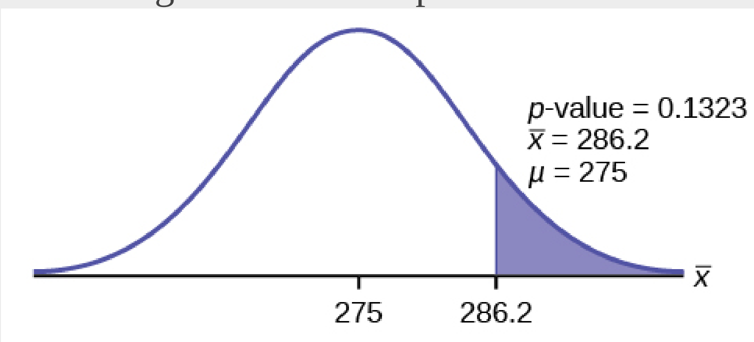
$x = 286.2$ pounds (from the data).

σ = 55 pounds **(Always use σ if you know it.)** We assume μ = 275 pounds unless our data shows us otherwise.

Calculate the *p*-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$p\text{-value} = P(x > 286.2) = 0.1323.$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.



Compare α and the *p*-value:

α = 0.025 *p*-value = 0.1323

Make a decision: Since α <*p*-value, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The *p*-value can easily be calculated.

Note:
Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where

you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to > μ_0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the *p*-value ($p = 0.1331$, a little different from the previous calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (*z*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (*p*-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example:
Exercise:

Problem:

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Solution:

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

H_0: $\mu = 65$ H_a: $\mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: X = average score on the first statistics test.

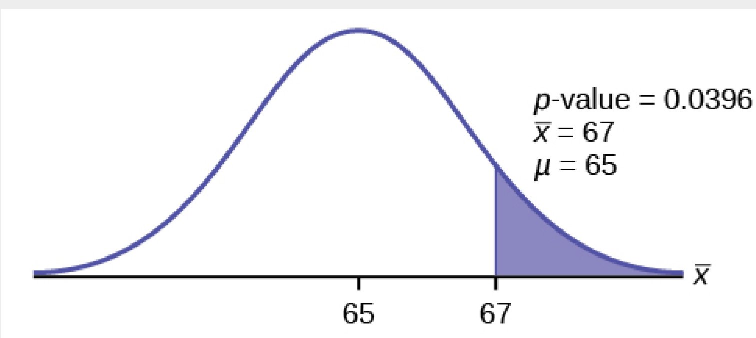
Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given n = 10 sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where n = 10 and df = 10 - 1 = 9.

Calculate the p-value using the Student's t-distribution:

p-value = $P(x > 67)$ = 0.0396 where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the *p*-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.



Compare α and the p-value:

Since α = 0.05 and p-value = 0.0396. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The *p*-value can easily be calculated.

Note:
Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the *p*-value ($p = 0.0396$) but it also calculates the test statistic (*t*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (*p*-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Note:
Try It
Exercise:

Problem:

It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

Solution:

H_0: $\mu = 5$

H_a: $\mu < 5$

$p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than \$5 a week.

Type I Error: To conclude that the stock price is growing slower than \$5 a week when, in fact, the stock price is growing at \$5 a week (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that the stock price is growing at a rate of \$5 a week when, in fact, the stock price is growing slower than \$5 a week (do not reject the null hypothesis when the null hypothesis is false).

Example:
Exercise:

Problem:

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution:

Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

H_0: $p = 0.50$ H_a: $p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ Therefore, $P' \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$

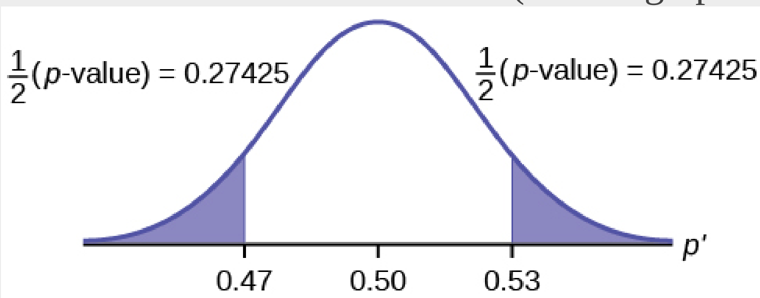
where $p = 0.50$, $q = 1-p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

p-value $= P(p' < 0.47 \text{ or } p' > 0.53) = 0.5485$

where $x = 53$, $p' = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the *p*-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in [link]).



$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 – 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the *p*-value:

Since $\alpha = 0.01$ and *p*-value = 0.5485. $\alpha <$ *p*-value.

Make a decision: Since $\alpha <$ *p*-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The *p*-value can easily be calculated.

Note:
Press `STAT` and arrow over to `TESTS`. Press `5:1-PropZTest`. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator calculates the *p*-value

(p = 0.5485) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with z = 0.6 (test statistic) and p = 0.5485 (p-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Note:
Try It
Exercise:

Problem:

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

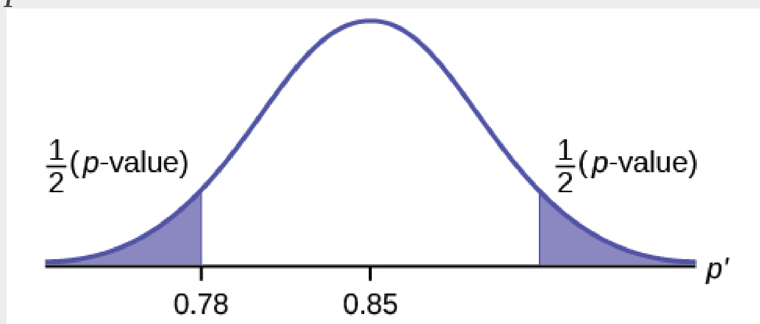
Solution:

Since the problem is about percentages, this is a test of single population proportions.

$H_0 : p = 0.85$

$H_a : p \neq 0.85$

$p = 0.7554$



Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example:
Exercise:

Problem:

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Solution:

Set up the Hypothesis Test:

H_0: $p = 0.30$ H_a: $p \neq 0.30$

Determine the distribution needed:

The **random variable** is P' = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' \sim N\left(0.30, \sqrt{\frac{(0.30)\cdot(0.70)}{150}}\right)$

Exercise:

Problem:

a. The value that helps determine the p-value is p'. Calculate p'.

Solution:

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43$, $n = 150$

$p' = \frac{43}{150}$

Exercise:

Problem: b. What is a **success** for this problem?

Solution:

b. A success is having three cell phones in a household.

Exercise:

Problem: c. What is the level of significance?

Solution:

c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

Exercise:

Problem:

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.
Calculate the p-value.

Solution:

d. p-value = 0.7216

Exercise:

Problem:

e. Make a decision. ______________(Reject/Do not reject) H_0 because______________.

Solution:

e. Assuming that $\alpha = 0.05$, $\alpha < p$-value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Note:

Try It
Exercise:

Problem:

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

Solution:

H_0: $p = 0.92$

H_a: $p < 0.92$

p-value = 0.0046

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the

null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example:
Exercise:

My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out

Problem: Use the new shampoo or go without?

Solution:

Set up the hypothesis test:

H_0: $p \leq 0.25$ H_a: $p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable X or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: $z = 2.3163$

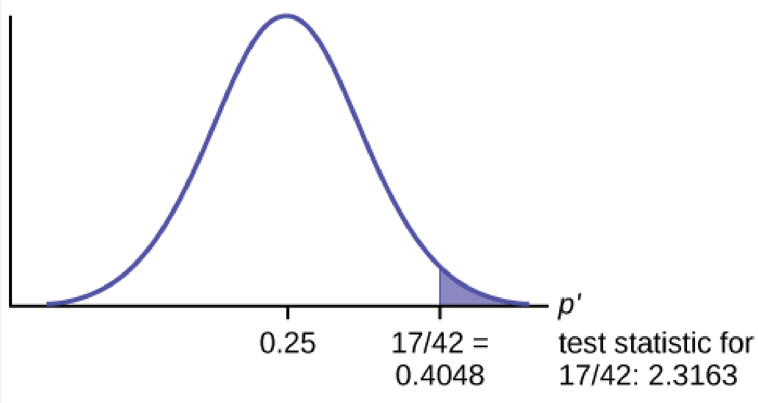
Calculate the p-value using the normal distribution for proportions:

p-value = 0.0103

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



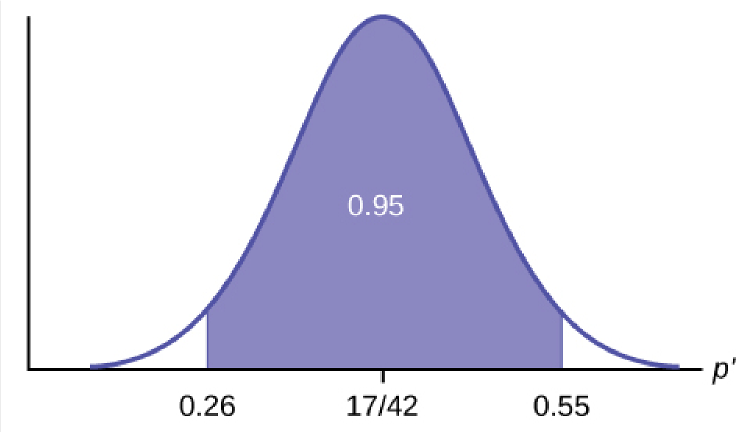
Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha < p$-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.



Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion *p* of fleas that are killed by the new shampoo is between 26% and 55%.

Note:
Note
This test result is not very definitive since the *p*-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example:
Exercise:

Problem:

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95
Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

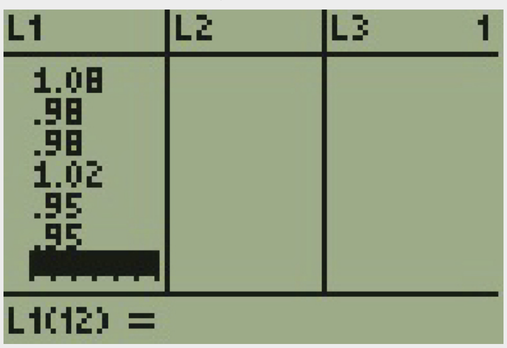
Solution:

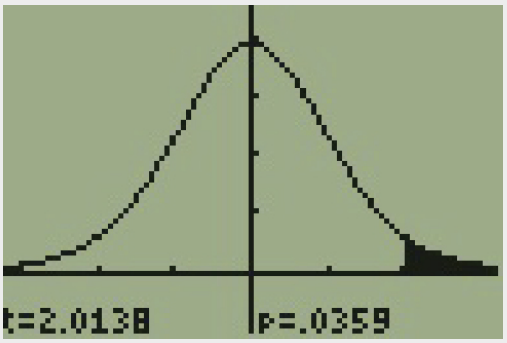
Let’s follow a four-step process to answer this statistical question.

1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be

    a. H_0: $\mu \leq 1$
    b. H_a: $\mu > 1$

2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.





4. **State the Conclusions**: Since the *p*-value* (p = 0.036) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example:
Exercise:

Problem:

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

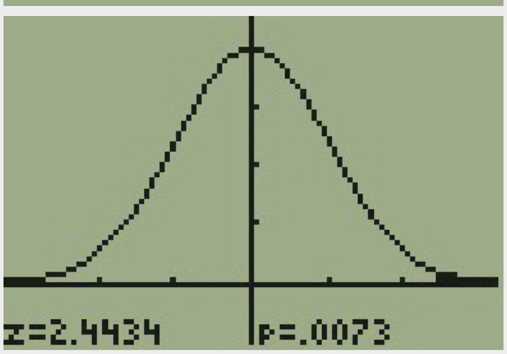
Solution:

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be

    a. H_0: $p \leq 0.00034$
    b. H_a: $p > 0.00034$

    If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420{,}019$. The sample is sufficiently large because we have $np = 420{,}019(0.00034) = 142.8$, $nq = 420{,}019(0.99966) = 419{,}876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are



4. Since the p-value = 0.0073 is greater than our alpha value = 0.005, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example:
Exercise:

Problem:

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

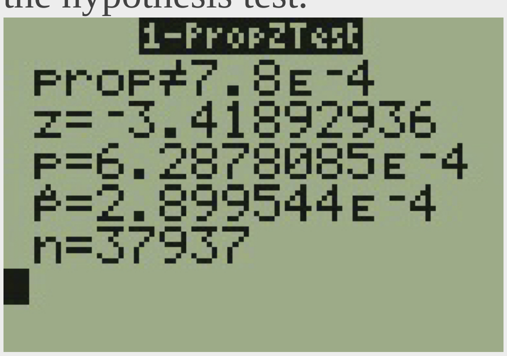
Solution:

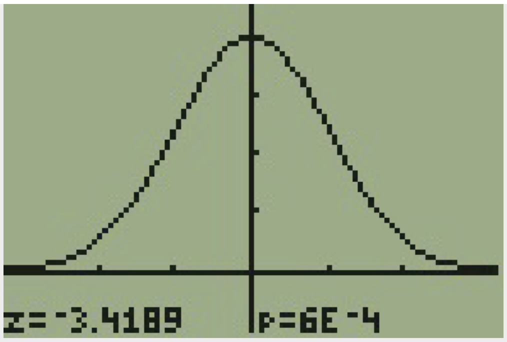
We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion *z*-test. The hypotheses for the test will be

   a. H_0: $p = 0.00078$
   b. H_a: $p \neq 0.00078$

3. The following screen shots display the summary statistics from the hypothesis test.





4. Since the *p*-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

Chapter Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the *p*-value. (A *z*-score and a *t*-score are examples of test statistics.)
5. Compare the preconceived α with the *p*-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the *p*-value. Remember that the quantity $1 – \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If

the power is low, the null hypothesis might not be rejected when it should be.

Exercise:

Problem:

Assume H_0: $\mu = 9$ and H_a: $\mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a left-tailed test.

Exercise:

Problem:

Assume H_0: $\mu \leq 6$ and H_a: $\mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise:

Problem:

Assume H_0: $p = 0.25$ and H_a: $p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a two-tailed test.

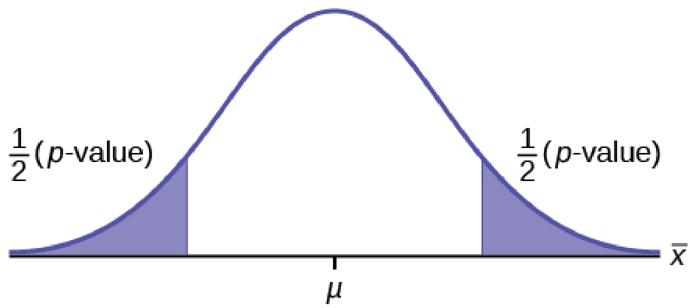
Exercise:

Problem: Draw the general graph of a left-tailed test.

Exercise:

Problem: Draw the graph of a two-tailed test.

Solution:



Exercise:

Problem:

A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Exercise:

Problem:

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

Solution:

a right-tailed test

Exercise:

Problem:

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

Exercise:

Problem:

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

Solution:

a left-tailed test

Exercise:

Problem:

If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

Exercise:

Problem:

Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a left-tailed test.

Exercise:

Problem:

Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise:

Problem:

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a two-tailed test.

Homework

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note:
Note
If you are using a Student's-*t* distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

Exercise:

Problem:

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, is the data highly inconsistent with the claim?

Solution:

a. H_0: $\mu \geq 50{,}000$
b. H_a: $\mu < 50{,}000$
c. Let $\bar{X}$ = the average lifespan of a brand of tires.
d. normal distribution
e. z = -2.315
f. p-value = 0.0103
g. Check student's solution.

h.
   i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.

i. (43,537, 49,463)

Exercise:

Problem:

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

Exercise:

Problem:

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

Solution:

a. H_0: μ = $1.00
b. H_a: $\mu \neq$ $1.00
c. Let $\bar{X}$ = the average cost of a daily newspaper.
d. normal distribution
e. $z = -0.866$
f. *p*-value = 0.3865
g. Check student's solution.

h.
   i. Alpha: 0.01
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.01.
   iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $1. The mean cost could be $1.

i. ($0.84, $1.06)

Exercise:

Problem:

An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

Exercise:

Problem:

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?

Solution:

a. H_0: $\mu = 10$
b. H_a: $\mu \neq 10$
c. Let $\bar{X}$ the mean number of sick days an employee takes per year.
d. Student's t-distribution
e. $t = –1.12$
f. p-value = 0.300

g. Check student’s solution.

h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The *p*-value is greater than 0.05.
iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.

i. (4.9443, 11.806)

Exercise:

Problem:

In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

Exercise:

Problem:

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

Solution:

a. H_0: $p \geq 0.6$
b. H_a: $p < 0.6$
c. Let P' = the proportion of students who feel more enriched as a result of taking Elementary Statistics.
d. normal for a single proportion
e. 1.12
f. p-value = 0.1308
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. Confidence Interval: (0.409, 0.654)
The "plus-4s" confidence interval is (0.411, 0.648)

Exercise:

Problem:

A Nissan Motor Corporation advertisement read, "The average man's I.Q. is 107. The average brown trout's I.Q. is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

Exercise:

Problem:

Refer to Exercise 9.119. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout's mean I.Q. is **not** four.

Solution:

a. H_0: $\mu = 4$
b. H_a: $\mu \neq 4$
c. Let X the average I.Q. of a set of brown trout.
d. two-tailed Student's t-test
e. $t = 1.95$
f. p-value = 0.076
g. Check student’s solution.

h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05
iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

i. (3.8865,5.9468)

Exercise:

Problem:

According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don’t believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

Exercise:

Problem:

A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent’s survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

Solution:

a. H_0: $p \geq 0.13$
b. H_a: $p < 0.13$
c. Let P' = the proportion of Americans who have seen or sensed angels
d. normal for a single proportion
e. –2.688
f. p-value = 0.0036
g. Check student's solution.

h.
   i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.

i. (0, 0.0623).
The“plus-4s” confidence interval is (0.0022, 0.0978)

Exercise:

Problem:

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Exercise:

Problem:

Use the “Lap time” data for Lap 4 (see [link]) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

Solution:

a. H_0: $\mu \geq 129$
b. H_a: $\mu < 129$
c. Let $\bar{X}$ = the average time in seconds that Terri finishes Lap 4.
d. Student's *t*-distribution
e. $t = 1.209$
f. 0.8792
g. Check student’s solution.
h. i. Alpha: 0.05
 ii. Decision: Do not reject the null hypothesis.
 iii. Reason for decision: The *p*-value is greater than 0.05.
 iv. Conclusion: There is insufficient evidence to conclude that Terri’s mean lap time is less than 129 seconds.
i. (128.63, 130.37)

Exercise:

Problem:

Use the “Initial Public Offering” data (see [link]) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note:
Note

The following questions were written by past students. They are excellent problems!

Exercise:

Problem: "Asian Family Reunion," by Chau Nguyen

Every two years it comes around.

We all get together from different towns.

In my honest opinion,

It's not a typical family reunion.

Not forty, or fifty, or sixty,

But how about seventy companions!

The kids would play, scream, and shout

One minute they're happy, another they'll pout.

The teenagers would look, stare, and compare

From how they look to what they wear.

The men would chat about their business

That they make more, but never less.

Money is always their subject

And there's always talk of more new projects.

The women get tired from all of the chats

They head to the kitchen to set out the mats.

Some would sit and some would stand

Eating and talking with plates in their hands.

Then come the games and the songs

And suddenly, everyone gets along!

With all that laughter, it's sad to say

That it always ends in the same old way.

They hug and kiss and say "good-bye"

And then they all begin to cry!

I say that 60 percent shed their tears

But my mom counted 35 people this year.

She said that boys and men will always have their pride,

So we won't ever see them cry.

I myself don't think she's correct,

So could you please try this problem to see if you object?

Solution:

a. H_0: $p = 0.60$
b. H_a: $p < 0.60$
c. Let P' = the proportion of family members who shed tears at a reunion.
d. normal for a single proportion
e. –1.71
f. 0.0438

g. Check student’s solution.

h. i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the p-value and alpha are quite close, so other tests should be done.

i. We are 95% confident that between 38.29% and 61.71% of family members will shed tears at a family reunion. (0.3829, 0.6171). The“plus-4s” confidence interval (see chapter 8) is (0.3861, 0.6139)

Note that here the “large-sample” 1 – PropZTest provides the approximate p-value of 0.0438. Whenever a p-value based on a normal approximation is close to the level of significance, the exact p-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

Exercise:

Problem: "The Problem with Angels," by Cyndy Dowling

Although this problem is wholly mine,

The catalyst came from the magazine, Time.

On the magazine cover I did find

The realm of angels tickling my mind.

Inside, 69% I found to be

In angels, Americans do believe.

Then, it was time to rise to the task,

Ninety-five high school and college students I did ask.

Viewing all as one group,

Random sampling to get the scoop.

So, I asked each to be true,

"Do you believe in angels?" Tell me, do!

Hypothesizing at the start,

Totally believing in my heart

That the proportion who said yes

Would be equal on this test.

Lo and behold, seventy-three did arrive,

Out of the sample of ninety-five.

Now your job has just begun,

Solve this problem and have some fun.

Exercise:

Problem: "Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,

I had to find some sane defense.

Some light and lifting simple play

To float my math anxiety away.

Blowing bubbles lifts me high

Takes my troubles to the sky.

POIK! They're gone, with all my stress

Bubble therapy is the best.

The label said each time I blew

The average number of bubbles would be at least 22.

I blew and blew and this I found

From 64 blows, they all are round!

But the number of bubbles in 64 blows

Varied widely, this I know.

20 per blow became the mean

They deviated by 6, and not 16.

From counting bubbles, I sure did relax

But now I give to you your task.

Was 22 a reasonable guess?

Find the answer and pass this test!

Solution:

a. H_0: $\mu \geq 22$
b. H_a: $\mu < 22$
c. Let $\bar{X}$ = the mean number of bubbles per blow.
d. Student's t-distribution
e. –2.667
f. p-value = 0.00486
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.

i. (18.501, 21.499)

Exercise:

Problem: "Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles

Bred puppies with numerous freckles

The Dalmatians he sought

Possessed spot upon spot

The more spots, he thought, the more shekels.

His competitors did not agree

That freckles would increase the fee.

They said, “Spots are quite nice

But they don't affect price;

One should breed for improved pedigree.”

The breeders decided to prove

This strategy was a wrong move.

Breeding only for spots

Would wreak havoc, they thought.
His theory they want to disprove.
They proposed a contest to Spreckles
Comparing dog prices to freckles.
In records they looked up
One hundred one pups:
Dalmatians that fetched the most shekels.
They asked Mr. Spreckles to name
An average spot count he'd claim
To bring in big bucks.
Said Spreckles, “Well, shucks,
It's for one hundred one that I aim.”
Said an amateur statistician
Who wanted to help with this mission.
“Twenty-one for the sample
Standard deviation's ample:
They examined one hundred and one
Dalmatians that fetched a good sum.
They counted each spot,
Mark, freckle and dot

And tallied up every one.

Instead of one hundred one spots

They averaged ninety six dots

Can they muzzle Spreckles'

Obsession with freckles

Based on all the dog data they've got?

Exercise:

Problem:

"Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:

Price per box of Mac and Cheese:

- 5 stores @ $2.02
- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27

- 7 stores @ \$1.50
- 5 stores @ \$1.89
- 8 stores @ 0.75.

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most \$1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

Solution:

a. H_0: $\mu \leq 1$
b. H_a: $\mu > 1$
c. Let X = the mean cost in dollars of macaroni and cheese in a certain town.
d. Student's *t*-distribution
e. $t = 0.340$
f. *p*-value = 0.36756
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.05
   iv. Conclusion: The mean cost could be \$1, or less. At the 5% significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than \$1.
i. (0.8291, 1.241)

Exercise:

Problem:

"William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi

THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet’s tutor
- HOROTIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter’s recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how cans’t thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn’d, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, “What piece of work is man, how noble in reason, how infinite in faculties? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to’t I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether ‘tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

Exercise:

Problem: "Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known

problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

Solution:

a. H_0: $p = 0.01$
b. H_a: $p > 0.01$
c. Let P' = the proportion of errors generated
d. Normal for a single proportion
e. 2.13
f. 0.0165
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.

i. Confidence interval: (0, 0.094).
The"plus-4s" confidence interval is (0.004, 0.144).

Exercise:

Problem: "Japanese Girls' Names"

by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

Exercise:

Problem: "Phillip’s Wish," by Suzanne Osorio

My nephew likes to play

Chasing the girls makes his day.

He asked his mother

If it is okay

To get his ear pierced.

She said, “No way!”

To poke a hole through your ear,

Is not what I want for you, dear.

He argued his point quite well,

Says even my macho pal, Mel,

Has gotten this done.

It’s all just for fun.

C’mon please, mom, please, what the hell.

Again Phillip complained to his mother,

Saying half his friends (including their brothers)

Are piercing their ears

And they have no fears

He wants to be like the others.

She said, “I think it’s much less.

We must do a hypothesis test.

And if you are right,

I won’t put up a fight.

But, if not, then my case will rest.”

We proceeded to call fifty guys

To see whose prediction would fly.

Nineteen of the fifty

Said piercing was nifty

And earrings they’d occasionally buy.

Then there’s the other thirty-one,

Who said they’d never have this done.

So now this poem’s finished.

Will his hopes be diminished,

Or will my nephew have his fun?

Solution:

a. H_0: p = 0.50
b. H_a: p < 0.50
c. Let P' = the proportion of friends that has a pierced ear.
d. normal for a single proportion
e. –1.70
f. p-value = 0.0448
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The *p*-value is less than 0.05. (However, they are very close.)
   iv. Conclusion: There is sufficient evidence to support the claim that less than 50% of his friends have pierced ears.

i. Confidence Interval: (0.245, 0.515): The “plus-4s” confidence interval is (0.259, 0.519).

Exercise:

Problem: "The Craven," by Mark Salangsang

Once upon a morning dreary

In stats class I was weak and weary.

Pondering over last night’s homework

Whose answers were now on the board

This I did and nothing more.

While I nodded nearly napping

Suddenly, there came a tapping.

As someone gently rapping,

Rapping my head as I snore.

Quoth the teacher, “Sleep no more.”

“In every class you fall asleep,”

The teacher said, his voice was deep.

“So a tally I’ve begun to keep

Of every class you nap and snore.

The percentage being forty-four.”

“My dear teacher I must confess,

While sleeping is what I do best.

The percentage, I think, must be less,

A percentage less than forty-four.”

This I said and nothing more.

“We’ll see,” he said and walked away,

And fifty classes from that day

He counted till the month of May

The classes in which I napped and snored.

The number he found was twenty-four.

At a significance level of 0.05,

Please tell me am I still alive?

Or did my grade just take a dive

Plunging down beneath the floor?

Upon thee I hereby implore.

Exercise:

Problem:

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

Solution:

a. H_0: $p = 0.40$
b. H_a: $p < 0.40$
c. Let P' = the proportion of schoolmates who fear public speaking.
d. normal for a single proportion
e. –1.01
f. p-value = 0.1563
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.

i. Confidence Interval: (0.3241, 0.4240): The "plus-4s" confidence interval is (0.3257, 0.4250).

Exercise:

Problem:

Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

Exercise:

Problem:

According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

Solution:

a. H_0: $p = 0.14$
b. H_a: $p < 0.14$
c. Let P' = the proportion of NYC residents that smoke.
d. normal for a single proportion
e. –0.2756
f. p-value = 0.3914
g. Check student's solution.

h.
   i. alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is

less than 0.14.

i. Confidence Interval: (0.0502, 0.2070): The "plus-4s" confidence interval (see chapter 8) is (0.0676, 0.2297).

Exercise:

Problem:

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

Exercise:

Problem:

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

Solution:

a. H_0: $\mu = 69{,}110$
b. H_a: $\mu > 69{,}110$
c. Let X = the mean salary in dollars for California registered nurses.
d. Student's *t*-distribution
e. $t = 1.719$
f. *p*-value: 0.0466
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is less than 0.05.

iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

i. ($68,757, $73,485)

Exercise:

Problem:

La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

Exercise:

Problem:

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?
After conducting the test, your decision and conclusion are

a. Reject H_0: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject H_0: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject H_0: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.

d. Reject H_0: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

Solution:

c

Exercise:

Problem:

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

Exercise:

Problem:

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

Solution:

c

Instructions: For the following ten exercises,
Hypothesis testing: For the following ten exercises, answer each question.

a. State the null and alternate hypothesis.
b. State the *p*-value.
c. State alpha.
d. What is your decision?
e. Write a conclusion.
f. Answer any other questions asked in the problem.

Exercise:

Problem:

According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

Exercise:

Problem:

A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

Solution:

a. H_0: $p = 0.488$ H_a: $p \neq 0.488$
b. p-value = 0.0114
c. alpha = 0.05
d. Reject the null hypothesis.
e. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.
f. The survey does not appear to be accurate.

Exercise:

Problem:

Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using α = 0.05, is the AAA proportion accurate?

Exercise:

Problem:

The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

Solution:

a. H_0: $p = 0.517$ H_a: $p \neq 0.517$
b. p-value = 0.9203.
c. alpha = 0.05.
d. Do not reject the null hypothesis.
e. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

Exercise:

Problem:

For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

Exercise:

Problem:

The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

Solution:

a. H_0: $\mu \geq 11.52$ H_a: $\mu < 11.52$
b. *p*-value = 0.000002 which is almost 0.
c. alpha = 0.05.
d. Reject the null hypothesis.
e. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1% because the *p*-value is almost 0.

Exercise:

Problem:

A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

Exercise:

Problem:

A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3 2 1 3 7 2 9 4 6 6 8 0 5 6 4 2 1 3 4 1

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?

Solution:

a. H_0: $\mu \leq 5.8$ H_a: $\mu > 5.8$
b. p-value = 0.9987
c. alpha = 0.05
d. Do not reject the null hypothesis.
e. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

Exercise:

Problem:

According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:
5 4 5 4 4 3 6 4 3 3 5 5 6 3 3 2 7 4 5 2 2 2 3 2
At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

Exercise:

Problem:

The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

Solution:

a. H_0: $\mu \geq 150$ H_a: $\mu < 150$
b. p-value = 0.0622
c. alpha = 0.01
d. Do not reject the null hypothesis.
e. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

References

Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.

Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.

Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).

Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).

Data from *Growing by Degrees* by Allen and Seaman.

Data from La Leche League International. Available online at http://www.lalecheleague.org/Law/BAFeb01.html.

Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).

Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).

Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

Data from the U.S. Census Bureau, available online at http://quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).

Data from the United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/.

Data from Toastmasters International. Available online at http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1.

Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).

Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).

“Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf.

Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).

Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at http://www.rainn.org/get-information/statistics/frequency-of-sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size *n* and we are interested in two new RVs - the sample mean, X, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size *n* of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal *n* times the population mean. The standard deviation

of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Introduction
class="introduction"

Linear regression and correlation can help you determine if an auto mechanic’s salary is related to his work experience. (credit: Joshua Rothhaas)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

Equation:

$$y = a + \mathrm{bx}$$

where *a* and *b* are constant numbers.

The variable ***x* is the independent variable, and *y* is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example:

The following examples are linear equations.

Equation:

$$y = 3 + 2\mathrm{x}$$

Equation:

$$y = –0.01 + 1.2\mathrm{x}$$

Note:

Try It

Exercise:

Problem: Is the following an example of a linear equation?

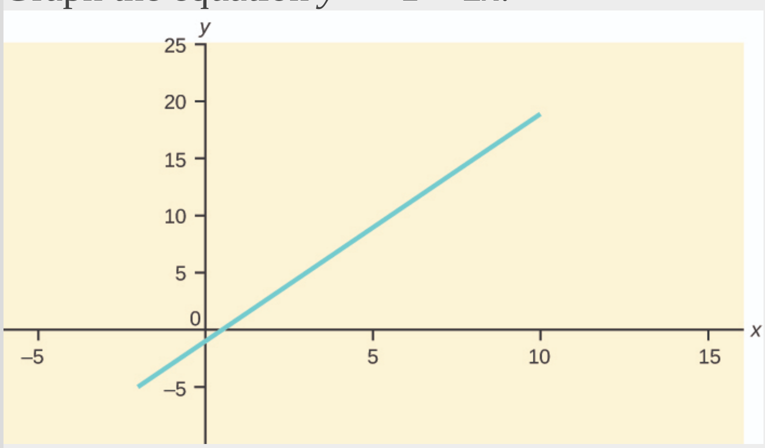
$y = –0.125 – 3.5x$

Solution:

yes

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

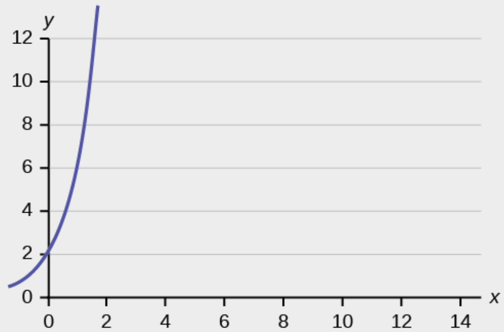
Example:
Graph the equation $y = –1 + 2x$.



Note:
Try It
Exercise:

Problem:

Is the following an example of a linear equation? Why or why not?



Solution:

No, the graph is not a straight line; therefore, it is not a linear equation.

Example:
Aaron's Word Processing Service (AWPS) does word processing. The rate for services is \$32 per hour plus a \$31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.
Exercise:

Problem:

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Solution:

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The \$31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Note:
Try It
Exercise:

Problem:

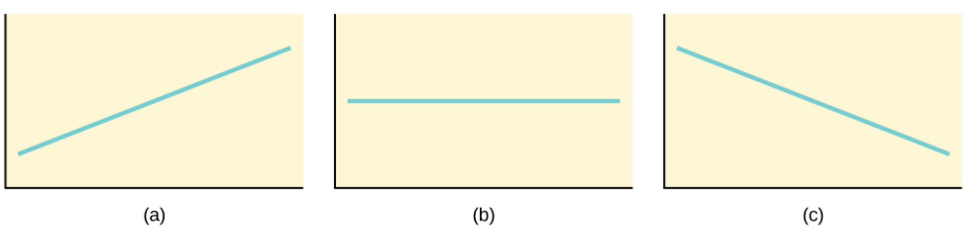
Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of \$50 per class as well as \$20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Solution:

$y = 50 + 20x$

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.



Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example:
Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.
Exercise:

Problem:

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Note:
Try It
Exercise:

Problem:

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns $20 for each hour he works.

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for agency reporting flu cases and TB Prevention.

Chapter Review

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $\boldsymbol{y = mx + b}$, where $\boldsymbol{m}$ and $\boldsymbol{b}$ are constants, $\boldsymbol{x}$ is the

independent variable, $\boldsymbol{y}$ is the dependent variable. In a statistical context, a linear equation is written in the form $\boldsymbol{y = a + bx}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the $\boldsymbol{x}$ variable (b is called a coefficient) is called as the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called as the y-intercept. Graphically, the y-intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The ***y*-intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

Formula Review

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of \$25 and another fee of \$12.50 an hour.

Exercise:

Problem: What are the dependent and independent variables?

Solution:

dependent variable: fee amount; independent variable: time

Exercise:

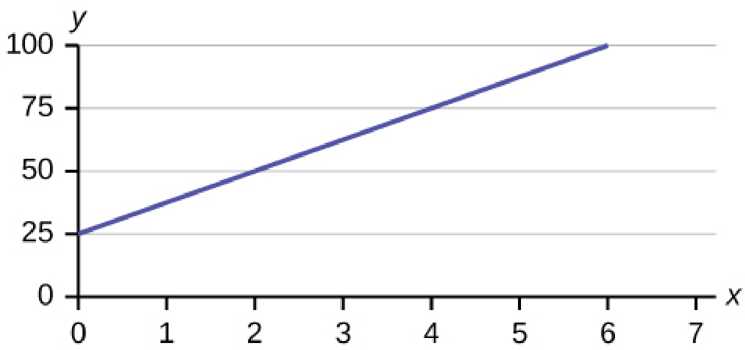
Problem:

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise:

Problem: Graph the equation from [link].

Solution:



Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise:

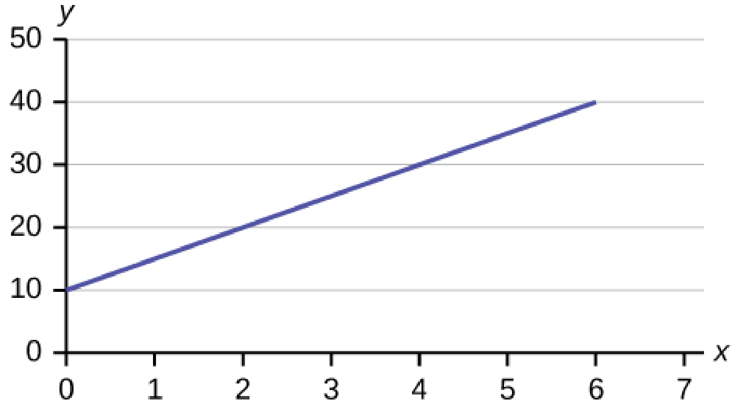
Problem:

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise:

Problem: Graph the equation from [link].

Solution:



Exercise:

Problem: Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise:

Problem: Which of the following equations are linear?

a. $y = 6x + 8$

b. $y + 7 = 3x$

c. $y - x = 8x^2$

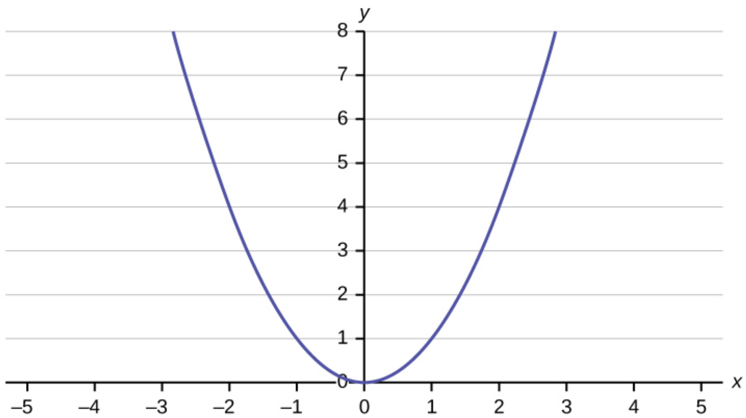
d. $4y = 8$

Solution:

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

Exercise:

Problem: Does the graph show a linear equation? Why or why not?



[link] contains real data for the first two decades of flu reporting.

Year	# flu cases diagnosed	# flu deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868

1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Adults and Adolescents only, United States

Exercise:

Problem:

Use the columns "year" and "# flu cases diagnosed. Why is “year” the independent variable and “# flu cases diagnosed.” the dependent variable (instead of the reverse)?

Solution:

The number of flu cases depends on the year. Therefore, year becomes the independent variable and the number of flu cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

Exercise:

Problem: What are the independent and dependent variables?

Exercise:

Problem:

What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

Use the following information to answer the next three questions. Due to

erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

Exercise:

Problem: What are the independent and dependent variables?

Exercise:

Problem: How many pounds of soil does the shoreline lose in a year?

Solution:

12,000 pounds of soil

Exercise:

Problem: What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise:

Problem: What are the slope and y-intercept? Interpret their meaning.

Solution:

The slope is –1.5 (b = –1.5). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is $15 ($a$ = 15). This means the price of stock before the trading day was $15.

Exercise:

Problem:

If you owned this stock, would you want a positive or negative slope? Why?

Homework

Exercise:

Problem:

For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

Solution:

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

Exercise:

Problem:

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	\$4,000 with an additional \$125 added per percentage point from 81–99%	\$6,500 with an additional \$125 added per percentage point from 101–119%	\$9,500 with an additional \$125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

Scatter Plots

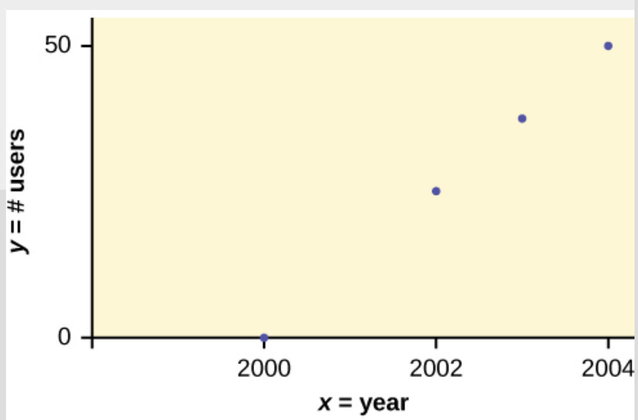
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

Example:
In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let *x* = the year and let *y* = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0

Scatter plot showing the number of m-commerce users (in millions) by year.



x (year)	y (# of users)
2004	47.0

Note: To create a scatter plot:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Note:
Try It
Exercise:

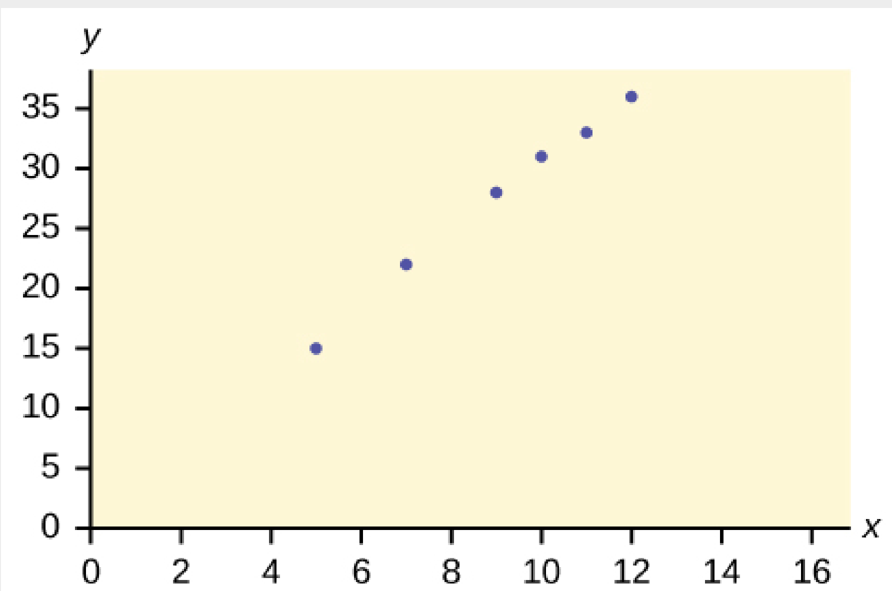
Problem:

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	*Y* (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Solution:



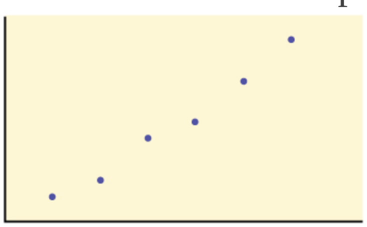
Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:

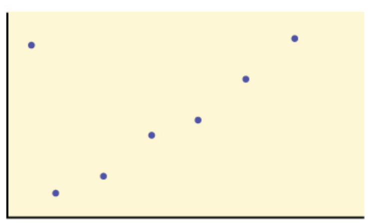
- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

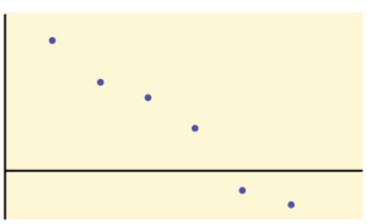
When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



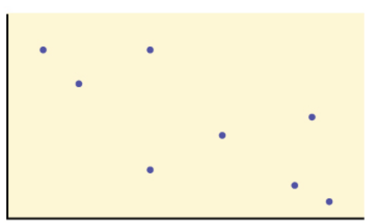
(a) Positive linear pattern (strong)



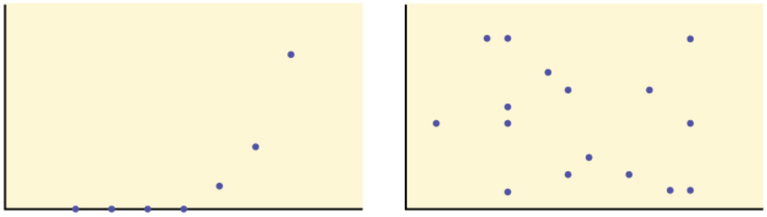
(b) Linear pattern w/ one deviation



(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern (b) No pattern

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

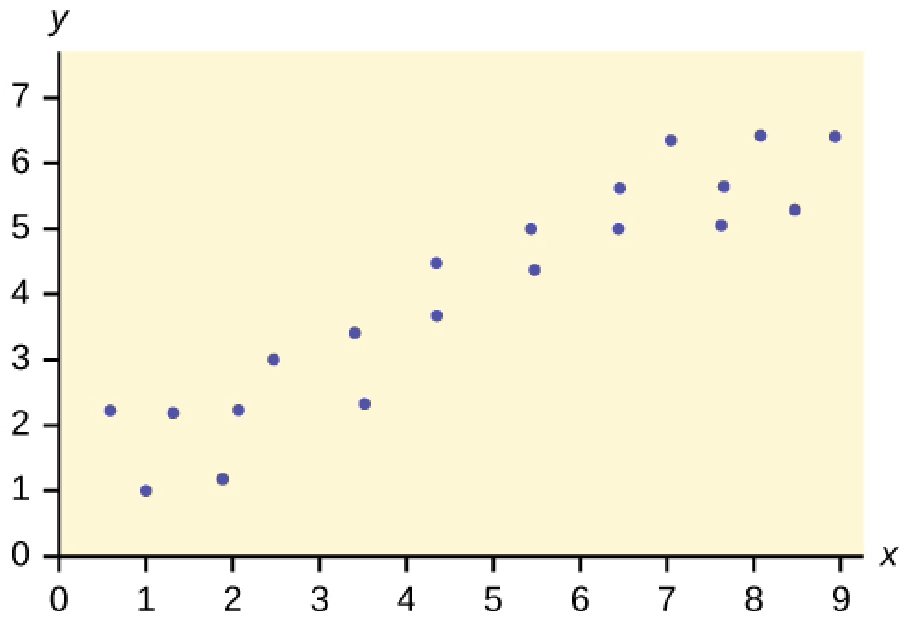
Chapter Review

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

Exercise:

Problem:

Does the scatter plot appear linear? Strong or weak? Positive or negative?



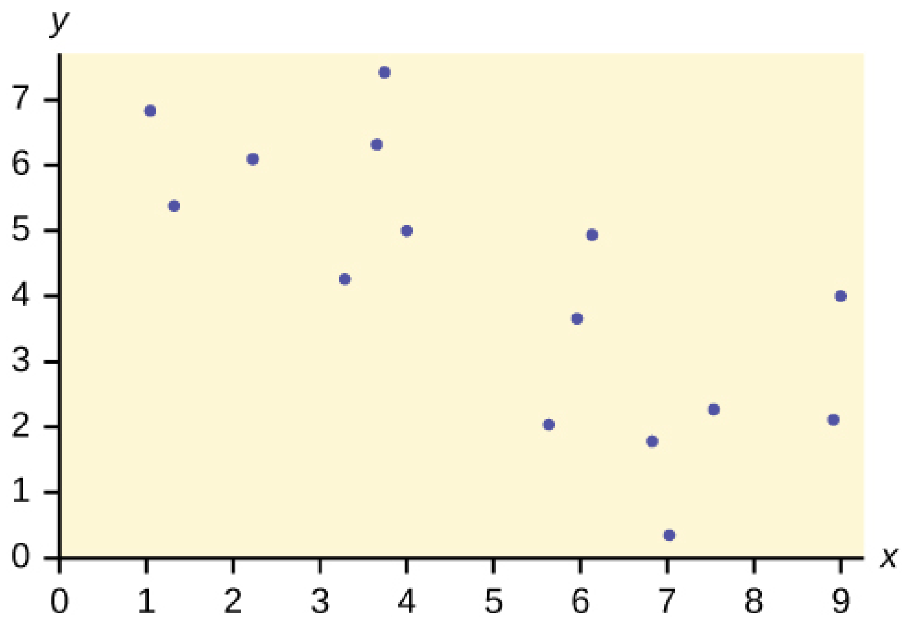
Solution:

The data appear to be linear with a strong, positive correlation.

Exercise:

Problem:

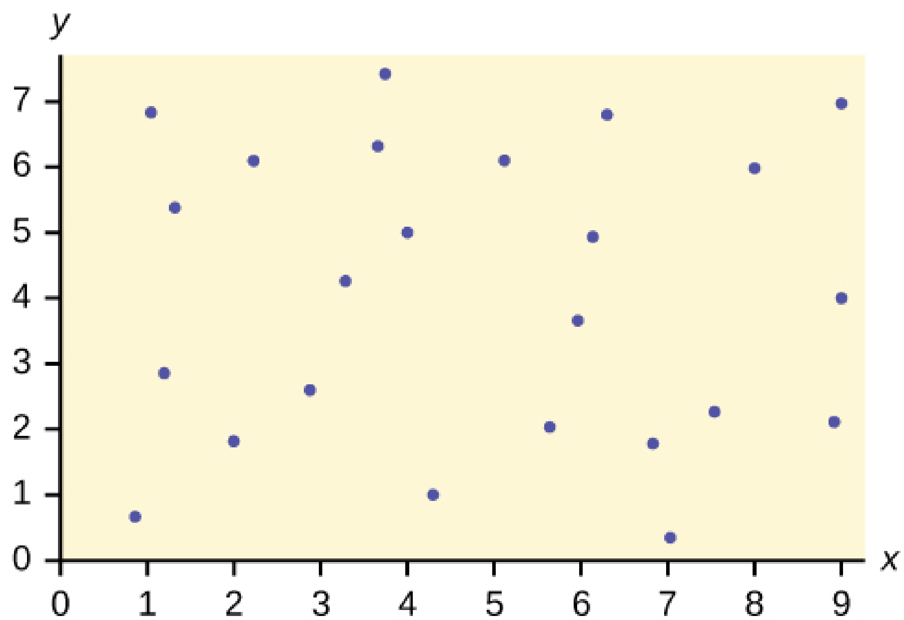
Does the scatter plot appear linear? Strong or weak? Positive or negative?



Exercise:

Problem:

Does the scatter plot appear linear? Strong or weak? Positive or negative?



Solution:

The data appear to have no correlation.

Homework

Exercise:

Problem:

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. [link] shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Solution:

Check student's solution.

Exercise:

Problem:

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Exercise:

Problem:

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Solution:

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

Exercise:

Problem:

If the level of significance is 0.05 and the *p*-value is 0.06, what conclusion can you draw?

Exercise:

Problem:

If there are 15 data points in a set of data, what is the number of degree of freedom?

Solution:

13

The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least-Squares Line**.

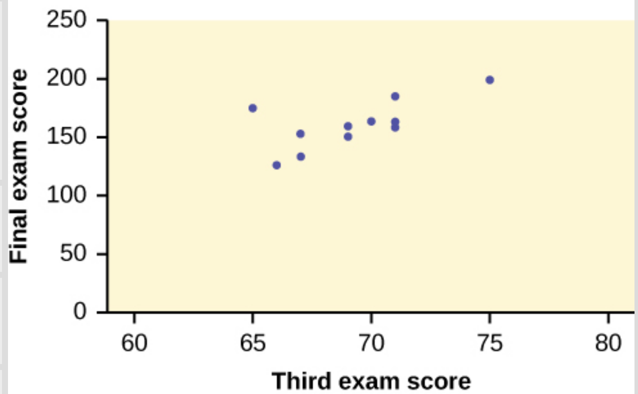
Note:
Collaborative Exercise
If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example:
A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table showing the scores on the final exam based on scores from the third exam.

Scatter plot showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159



Note:
Try It
Exercise:

Problem:

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in [link] show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

X **(depth in feet)**	***Y*** **(maximum dive time)**
50	80
60	55
70	45
80	35
90	25
100	22

Solution:

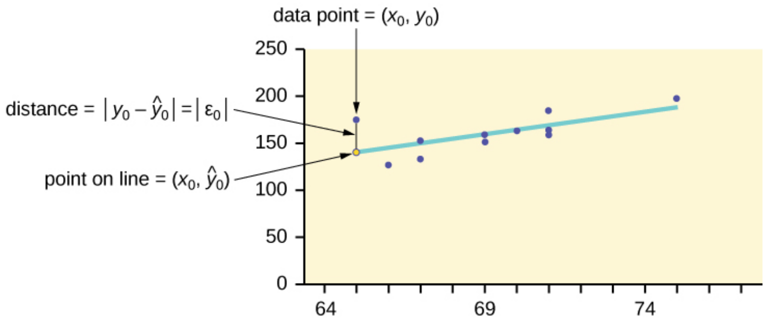
$\hat{y} = 127.24 – 1.11x$

At 110 feet, a diver could dive for only five minutes.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point ofthe line of best fit using least-squares linear regression has the form (x, $\hat{y}$).

The $\hat{y}$ is read **"*y* hat"** and is the **estimated value of *y***. It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram in [link], $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If yousquare each ε and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + ... + (\varepsilon_{11})^2 = \underset{i=1}{\overset{11}{\Sigma}} \varepsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:
Equation:

$$\hat{y} = a + bx$$

where $a = y - bx$ and $b = \frac{\Sigma(x-x)(y-y)}{\Sigma(x-x)^2}$.

The sample means of the x values and the y values are x and y, respectively. The best fit line always passes through the point (x, y).

The slope b can be written as $b = r\left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

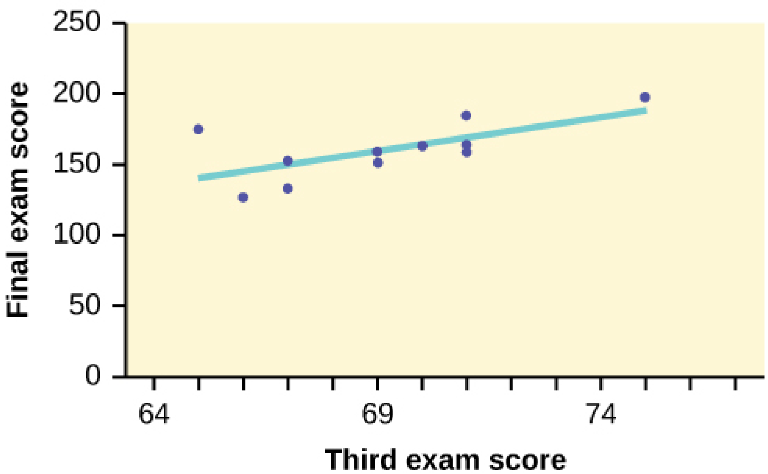
Least Squares Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are

scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line**
.

> **Note:**
> Note
> Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:
The graph of the line of best fit for the third-exam/final-exam example is as follows:



The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

Equation:

$$\hat{y} = -173.51 + 4.83x$$

> **Note:**
> Reminder
> Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for *y* given *x* within the domain of *x*-values in the sample data, **but not necessarily for *x*-values outside that domain.** You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the *x*-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

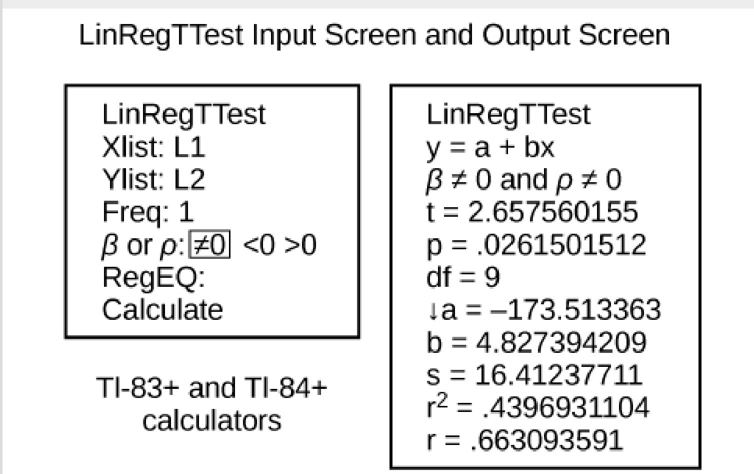
The slope of the line, *b*, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (*y*) changes for every one unit increase in the independent (*x*) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE
Slope: The slope of the line is $b = 4.83$.
Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Note:
Using the Linear Regression T Test: LinRegTTest

1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (*x*,*y*) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
3. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
4. On the next line, at the prompt β or ρ, highlight "≠ 0" and press ENTER
5. Leave the line for "RegEq:" blank
6. Highlight Calculate and press ENTER.



The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.
The second line says $y = a + bx$. Scroll down to find the values $a = –173.513$, and $b = 4.8273$; the equation of the best fit line is $\hat{y} = –173.51 + 4.83x$
The two items at the bottom are $r_2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.
Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1
3. On the input screen for PLOT 1, highlight **On**, and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
8. To graph the best-fit line, press the "Y=" key and type the equation –173.5 + 4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

Note:
NOTE
Another way to graph the line after you create a scatter plot is to use LinRegTTest.

1. Make sure you have done the scatter plot. Check it on your screen.
2. Go to LinRegTTest and enter the lists.
3. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
4. Press Y = (you will see the regression equation).
5. Press GRAPH. The line will be drawn."

The Correlation Coefficient *r*

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between *x* and *y*.

The **correlation coefficient, *r*,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable *x* and the dependent variable *y*.

The correlation coefficient is calculated as
Equation:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where *n* = the number of data points.

If you suspect a linear relationship between *x* and *y*, then *r* can measure how strong the linear relationship is.

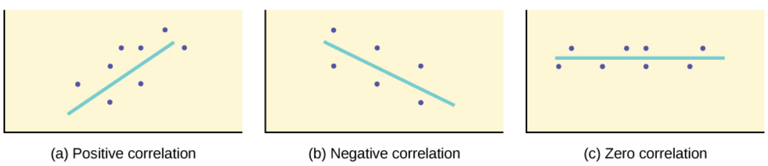
What the VALUE of *r* tells us:

- The value of *r* is always between –1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation *r* indicates the strength of the linear relationship between *x* and *y*. Values of *r* close to –1 or to +1 indicate a stronger linear relationship between *x* and *y*.
- If *r* = 0 there is absolutely no linear relationship between *x* and *y* **(no linear correlation)**.
- If *r* = 1, there is perfect positive correlation. If *r* = –1, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course,in the real world, this will not generally happen.

What the SIGN of *r* tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best-fit line.

Note:
Note
Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."**



(a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called the coefficient of determination and is the square of the correlation coefficient, but is usually stated as a percent, rather

than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable *y* that can be explained by variation in the independent (explanatory) variable *x* using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in *y* that is NOT explained by variation in *x* using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation (1 – 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Chapter Review

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the *x* and *y* variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of *y* and the estimated value of *y*. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Exercise:

Problem: Draw a scatter plot of the data.

Exercise:

Problem: Use regression to find the equation for the line of best fit.

Solution:

$\hat{y} = 2.23 + 1.99x$

Exercise:

Problem: Draw the line of best fit on the scatter plot.

Exercise:

Problem:

What is the slope of the line of best fit? What does it represent?

Solution:

The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

Exercise:

Problem:

What is the *y*-intercept of the line of best fit? What does it represent?

Exercise:

Problem: What does an r value of zero mean?

Solution:

It means that there is no correlation between the data sets.

Exercise:

Problem: When $n = 2$ and $r = 1$, are the data significant? Explain.

Exercise:

Problem:

When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

Solution:

Yes, there are enough data points and the value of r is strong enough to show that there is a strong negative correlation between the data sets.

Homework

Exercise:

Problem:

What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

Exercise:

Problem: Explain what it means when a correlation has an r^2 of 0.72.

Solution:

It means that 72% of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

Exercise:

Problem:

Can a coefficient of determination be negative? Why or why not?

Glossary

Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

Equation:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where *n* is the number of data points. The coefficient cannot be more then 1 and less then –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between *x* and *y*.

Introduction
class="introduction"

One-way ANOVA is used to measure information from several groups.

Note:

Chapter Objectives

By the end of this chapter, the student should be able to:

- Interpret the *F* probability distribution as the number of groups and the sample size change.
- Discuss two uses for the *F* distribution: one-way ANOVA and the test of two variances.
- Conduct and interpret one-way ANOVA.
- Conduct and interpret hypothesis tests of two variances.

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests comparing averages between more than two groups, statisticians have developed a method called "Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the *F* distribution, used for one-way ANOVA, and the test of two variances. This is just a very brief overview of one-way ANOVA. You will study this topic in much greater detail in future statistics courses. One-Way ANOVA, as it is presented here, relies heavily on a calculator or computer.

One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses **variances** to help determine if the means are equal or not. In order to perform a one-way ANOVA test, there are five basic **assumptions** to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have **equal standard deviations (or variances)**.
4. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

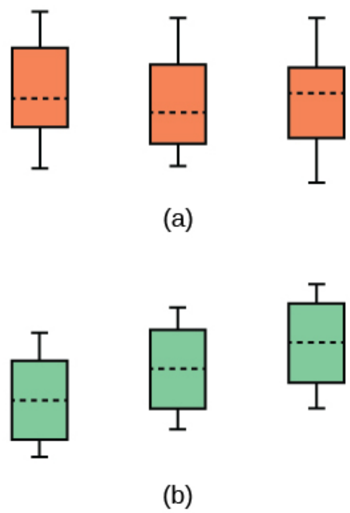
The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups:

H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

H_a: At least two of the group means $\mu_1, \mu_2, \mu_3, ..., \mu_k$ are not equal. That is, $\mu_i \neq \mu_j$ for some $i \neq j$.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), H_0: $\mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).



(a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

Chapter Review

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent, and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the *F* distribution with two different degrees of freedom.

Assumptions:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).

Use the following information to answer the next five exercises. There are five basic assumptions that must be fulfilled in order to perform a one-way ANOVA test. What are they?

Exercise:

Problem: Write one assumption.

Solution:

Each population from which a sample is taken is assumed to be normal.

Exercise:

Problem: Write another assumption.

Exercise:

Problem: Write a third assumption.

Solution:

The populations are assumed to have equal standard deviations (or variances).

Exercise:

Problem: Write a fourth assumption.

Exercise:

Problem: Write the final assumption.

Solution:

The response is a numerical value.

Exercise:

Problem:

State the null hypothesis for a one-way ANOVA test if there are four groups.

Exercise:

Problem:

State the alternative hypothesis for a one-way ANOVA test if there are three groups.

Solution:

H_a: At least two of the group means μ_1, μ_2, μ_3 are not equal.

Exercise:

Problem: When do you use an ANOVA test?

Homework

Exercise:

Problem:

Three different traffic routes are tested for mean driving time. The entries in the [link] are the driving times in minutes on the three different routes.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

State SS_{between}, SS_{within}, and the F statistic.

Solution:

$SS_{\text{between}} = 26$
$SS_{\text{within}} = 441$
$F = 0.2653$

Exercise:

Problem:

Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
	________	________	________	________	________
	________	________	________	________	________

State the hypotheses.

H_0: ____________

H_a: ____________

Glossary

Analysis of Variance
: also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

One-Way ANOVA
: a method of testing whether or not the means of three or more populations are equal; the method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.
- there is one independent variable and one dependent variable.

The test statistic for analysis of variance is the *F*-ratio.

Variance

mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as x – where x is a value of the data and

is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F* distribution**, named after Sir Ronald Fisher, an English statistician. The *F* statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if *F* follows an *F* distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

Note:
Note
The *F* distribution is derived from the Student's t-distribution. The values of the *F* distribution are squares of the corresponding values of the *t*-distribution. One-Way ANOVA expands the *t*-test for comparing more than two groups. The scope of that derivation is beyond the level of this course. It is preferable to use ANOVA when there are more than two groups instead of performing pairwise *t*-tests because performing multiple tests introduces the likelihood of making a Type 1 error.

To calculate the ***F* ratio**, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by *n* (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
2. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- SS_{between} = the **sum of squares** that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in Descriptive Statistics.

MS means "**mean square**." MS_{between} is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = total number of all the values combined (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x^2\right)}{n}$
- Total sum of squares: $\sum x^2 - \frac{\left(\sum x\right)^2}{n}$

- Explained variation: sum of squares representing variation among the different samples: $SS_{\text{between}} = \sum\left[\frac{(s_j)^2}{n_j}\right] - \frac{(\sum s_j)^2}{n}$
- Unexplained variation: sum of squares representing variation within samples due to chance: $SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$
- df's for different groups (df's for the numerator): $df = k - 1$
- Equation for errors within samples (df's for the denominator): $df_{\text{within}} = n - k$
- Mean square (variance estimate) explained by the different groups: $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$
- Mean square (variance estimate) that is due to chance (unexplained): $MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$

MS_{between} and MS_{within} can be written as follows:

- $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1}$
- $MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{SS_{within}}{n-k}$

The one-way ANOVA test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

Note:
Note
The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or F Statistic
$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}.Then the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The foregoing calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F-ratio can be written as:

F-Ratio Formula when the groups are the same size
$F = \frac{n \cdot s_{\bar{x}}{}^2}{s^2{}_{\text{pooled}}}$
where ...

- n = the sample size
- $df_{\text{numerator}} = k – 1$
- $df_{\text{denominator}} = n – k$
- s^2 pooled = the mean of the sample variances (pooled variance)
- $s_{\bar{x}}{}^2$ = the variance of the sample means

Data are typically put into a table for easy viewing. One-Way ANOVA results are often displayed in this manner by computer software.

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	*SS*(Factor)	$k – 1$	*MS*(Factor) = *SS*(Factor)/($k – 1$)	*F* = *MS*(Factor)/*MS*(Error)
Error (Within)	*SS*(Error)	$n – k$	*MS*(Error) = *SS*(Error)/($n – k$)	
Total	*SS*(Total)	$n – 1$		

Example:
Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in [link].

Plan 1: $n_1 = 4$	**Plan 2: $n_2 = 3$**	**Plan 3: $n_3 = 3$**
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

$s_1 = 16.5$, $s_2 = 15$, $s_3 = 15.5$
Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.
Equation:

$$SS(between) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n}$$

Equation:

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} - \frac{(s_1 + s_2 + s_3)^2}{10}$$

where $n_1 = 4$, $n_2 = 3$, $n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$

Equation:

$$= \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(15.5)^2}{3} - \frac{(16.5 + 15 + 15.5)^2}{10}$$

Equation:

$$SS(between) = 2.2458$$

Equation:

$$S(total) = \sum x^2 - \frac{(\sum x)^2}{n}$$

Equation:

$$= \left(5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2\right)$$

Equation:

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10}$$

Equation:

$$= 244 - \frac{47^2}{10} = 244 - 220.9$$

Equation:

$$SS(total) = 23.1$$

Equation:

$$SS(within) = SS(total) - SS(between)$$

Equation:

$$= 23.1 - 2.2458$$

Equation:

$$SS(within) = 20.8542$$

Note:
One-Way ANOVA Table: The formulas for *SS*(Total), *SS*(Factor) = *SS*(Between) and *SS*(Error) = *SS*(Within) as shown previously. The same information is provided by the TI calculator hypothesis test function ANOVA in STAT TESTS (syntax is ANOVA(L1, L2, L3) where L1, L2, L3 have the data from Plan 1, Plan 2, Plan 3 respectively).

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	*SS*(Factor) = *SS*(Between) = 2.2458	k – 1 = 3 groups – 1 = 2	*MS*(Factor) = *SS*(Factor)/(k – 1) = 2.2458/2 = 1.1229	F = *MS*(Factor)/*MS*(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	*SS*(Error) = *SS*(Within) = 20.8542	n – k = 10 total data – 3 groups = 7	*MS*(Error) = *SS*(Error)/(n – k) = 20.8542/7 = 2.9792	
Total	*SS*(Total) = 2.2458 + 20.8542 = 23.1	n – 1 = 10 total data – 1 = 9		

Note:
Try It
Exercise:

Problem:

As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments

- bare soil
- a commercial ground cover
- black plastic
- straw
- compost

All plants grew under the same conditions and were the same variety. Students recorded the weight (in grams) of tomatoes produced by each of the n = 15 plants:

Bare: n_1 = 3	**Ground Cover: n_2 = 3**	**Plastic: n_3 = 3**	**Straw: n_4 = 3**	**Compost: n_5 = 3**
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Create the one-way ANOVA table.

Solution:

Enter the data into lists L1, L2, L3, L4 and L5. Press STAT and arrow over to TESTS. Arrow down to ANOVA. Press ENTER and enter L1, L2, L3, L4, L5). Press ENTER. The table was filled in with the results from the calculator.

One-Way ANOVA table:

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	36,648,561	5 – 1 = 4	$\frac{36{,}648{,}561}{4} = 9,162,140$	$\frac{9{,}162{,}140}{2{,}044{,}672.6} = 4.4810$
Error (Within)	20,446,726	15 – 5 = 10	$\frac{20{,}446{,}726}{10} = 2,044,672.6$	
Total	57,095,287	15 – 1 = 14		

The one-way ANOVA hypothesis test is always right-tailed because larger *F*-values are way out in the right tail of the *F*-distribution curve and tend to make us reject H_0.

Notation

The notation for the *F* distribution is $F \sim F_{df(num),df(denom)}$

where $df(num) = df_{\text{between}}$ and $df(denom) = df_{\text{within}}$

The mean for the *F* distribution is $\mu = \frac{df(denom)}{df(denom)-2}$

References

Tomato Data, Marist College School of Science (unpublished student research)

Chapter Review

Analysis of variance compares the means of a response variable for several groups. ANOVA compares the variation within each group to the variation of the mean of each group. The ratio of these two is the *F* statistic from an *F* distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the ANOVA table.

Formula Review

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$df_{\text{between}} = df(num) = k - 1$

$df_{\text{within}} = df(denom) = n - k$

$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$

$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$

$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$

F ratio when the groups are the same size: $F = \frac{ns_{\bar{x}}{}^2}{s^2{}_{pooled}}$

Mean of the *F* distribution: $\mu = \frac{df(num)}{df(denom)-1}$

where:

- k = the number of groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group

- n = the total number of all values (observations) combined
- x = one value (one observation) from the data
- $s_{\overline{x}}{}^2$ = the variance of the sample means
- $s^2{}_{pooled}$ = the mean of the sample variances (pooled variance)

Use the following information to answer the next eight exercises. Groups of men from three different areas of the country are to be tested for mean weight. The entries in [link] are the weights for the different groups.

Group 1	**Group 2**	**Group 3**
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

Exercise:

Problem: What is the Sum of Squares Factor?

Solution:

4,939.2

Exercise:

Problem: What is the Sum of Squares Error?

Exercise:

Problem: What is the *df* for the numerator?

Solution:

2

Exercise:

Problem: What is the *df* for the denominator?

Exercise:

Problem: What is the Mean Square Factor?

Solution:

2,469.6

Exercise:

Problem: What is the Mean Square Error?

Exercise:

Problem: What is the *F* statistic?

Solution:

3.7416

Use the following information to answer the next eight exercises. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in the table are the goals per game for the different teams. The one-way ANOVA results are shown in [link].

Team 1	**Team 2**	**Team 3**	**Team 4**
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

Exercise:

Problem: What is $SS_{between}$?

Exercise:

Problem: What is the *df* for the numerator?

Solution:

3

Exercise:

Problem: What is $MS_{between}$?

Exercise:

Problem: What is SS_{within}?

Solution:

13.2

Exercise:

Problem: What is the *df* for the denominator?

Exercise:

Problem: What is MS_{within}?

Solution:

0.825

Exercise:

Problem: What is the *F* statistic?

Exercise:

Problem:

Judging by the *F* statistic, do you think it is likely or unlikely that you will reject the null hypothesis?

Solution:

Because a one-way ANOVA test is always right-tailed, a high *F* statistic corresponds to a low *p*-value, so it is likely that we will reject the null hypothesis.

Homework

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	**South**	**West**	**Central**	**East**
	16.3	16.9	16.4	16.2	17.1

	Northeast	South	West	Central	East
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\overline{x} =$	________	________	________	________	________
$s^2 =$	________	________	________	________	________

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H\alpha$: At least any two of the group means $\mu_1, \mu_2, \ldots, \mu_5$ are not equal.

Exercise:

Problem: degrees of freedom – numerator: $df(num)$ = _________

Exercise:

Problem: degrees of freedom – denominator: $df(denom)$ = ________

Solution:

$df(denom) = 15$

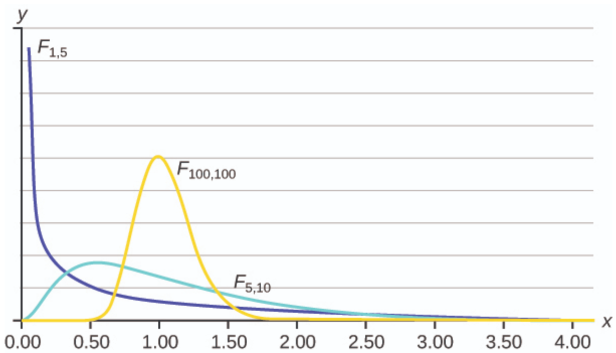
Exercise:

Problem: F statistic = ________

Facts About the F Distribution

Here are some facts about the *F* distribution.

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of *df*s.
3. The *F* statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
5. Other uses for the *F* distribution include comparing two variances and two-way Analysis of Variance. Two-Way Analysis is beyond the scope of this chapter.



Example:
Exercise:

Problem:

Let's return to the slicing tomato exercise in [link]. The means of the tomato yields under the five mulching conditions are represented by μ_1, μ_2, μ_3, μ_4, μ_5. We will conduct a hypothesis test to determine if all means are the same or at least one is different. Using a significance level of 5%, test the null hypothesis that there is no difference in mean yields among the five groups against the alternative hypothesis that at least one mean is different from the rest.

Solution:

The null and alternative hypotheses are:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: $\mu_i \neq \mu_j$ some $i \neq j$

The one-way ANOVA results are shown in [link]

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*

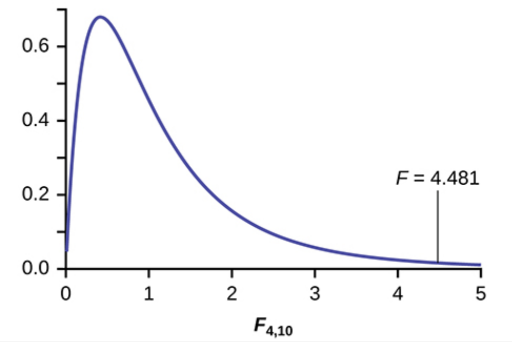
Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	36,648,561	5 – 1 = 4	$\frac{36{,}648{,}561}{4} = 9{,}162{,}140$	$\frac{9{,}162{,}140}{2{,}044{,}672.6} = 4.4810$
Error (Within)	20,446,726	15 – 5 = 10	$\frac{20{,}446{,}726}{10} = 2{,}044{,}672.6$	
Total	57,095,287	15 – 1 = 14		

Distribution for the test: $F_{4,10}$

$df(num) = 5 – 1 = 4$

$df(denom) = 15 – 5 = 10$

Test statistic: $F = 4.4810$



Probability Statement: p-value = $P(F > 4.481) = 0.0248$.

Compare α and the p-value: $\alpha = 0.05$, p-value = 0.0248

Make a decision: Since $\alpha > p$-value, we reject H_0.

Conclusion: At the 5% significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least some of mulches led to different mean yields.

Note:
To find these results on the calculator:
Press STAT. Press 1:EDIT. Put the data into the lists L_1, L_2, L_3, L_4, L_5.
Press STAT, and arrow over to TESTS, and arrow down to ANOVA. Press ENTER, and then enter L_1, L_2, L_3, L_4, L_5). Press ENTER. You will see that the values in the foregoing ANOVA table are easily produced by the calculator, including the test statistic and the p-value of the test.

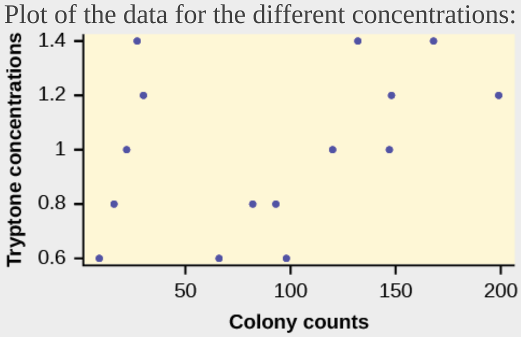
The calculator displays:
$F = 4.4810$
$p = 0.0248$ (p-value)
Factor
$df = 4$
$SS = 36648560.9$
$MS = 9162140.23$
Error
$df = 10$
$SS = 20446726$
$MS = 2044672.6$

Note:
Try It
Exercise:

Problem:

MRSA, or *Staphylococcus aureus*, can cause a serious bacterial infections in hospital patients. [link] shows various colony counts from different patients who may or may not have MRSA. The data from the table is plotted in Figure 13.5.

Conc = 0.6	**Conc = 0.8**	**Conc = 1.0**	**Conc = 1.2**	**Conc = 1.4**
9	16	22	30	27
66	93	147	199	168
98	82	120	148	132

Plot of the data for the different concentrations:


Test whether the mean number of colonies are the same or are different. Construct the ANOVA table (by hand or by using a TI-83, 83+, or 84+ calculator), find the p-value, and state your conclusion. Use a 5% significance level.

Solution:

While there are differences in the spreads between the groups (see [link]), the differences do not appear to be big enough to cause concern.

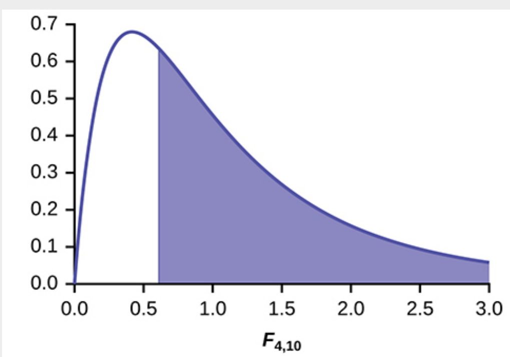
We test for the equality of mean number of colonies:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: $\mu^i \neq \mu^j$ some $i \neq j$

The one-way ANOVA table results are shown in [link].

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	10,233	5 – 1 = 4	$\frac{10{,}233}{4} = 2{,}558.25$	$\frac{2{,}558.25}{4{,}194.9} = 0.6099$
Error (Within)	41,949	15 – 5 = 10		
Total	52,182	15 – 1 = 14	$\frac{41{,}949}{10} = 4{,}194.9$	



Distribution for the test: $F_{4,10}$

Probability Statement: p-value = $P(F > 0.6099) = 0.6649$.

Compare α and the p-value: $\alpha = 0.05$, p-value = 0.669, $\alpha > p$-value

Make a decision: Since $\alpha > p$-value, we do not reject H_0.

Conclusion: At the 5% significance level, there is insufficient evidence from these data that different levels of tryptone will cause a significant difference in the mean number of bacterial colonies formed.

Example:

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in [link].

Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

MEAN GRADES FOR FOUR SORORITIES

Exercise:

Problem: Using a significance level of 1%, is there a difference in mean grades among the sororities?

Solution:

Let μ_1, μ_2, μ_3, μ_4 be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each five.

Note:
Note
This is an example of a **balanced design**, because each factor (i.e., sorority) has the same number of observations.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means μ_1, μ_2, μ_3, μ_4 are equal.

Distribution for the test: $F_{3,16}$

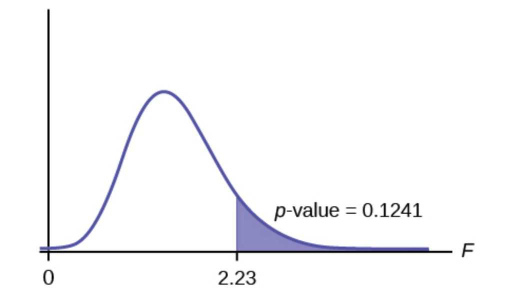
where $k = 4$ groups and $n = 20$ samples in total

$df(num) = k - 1 = 4 - 1 = 3$

$df(denom) = n - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph:



Probability statement: p-value = $P(F > 2.23) = 0.1241$

Compare α and the p-value: $\alpha = 0.01$
p-value = 0.1241
$\alpha < p$-value

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

Note:
Put the data into lists L_1, L_2, L_3, and L_4. Press `STAT` and arrow over to `TESTS`. Arrow down to `F:ANOVA`. Press `ENTER` and Enter (`L1,L2,L3,L4`).
The calculator displays the F statistic, the p-value and the values for the one-way ANOVA table:
$F = 2.2303$
$p = 0.1241$ (p-value)
Factor
$df = 3$
$SS = 2.88732$
$MS = 0.96244$
Error
$df = 16$
$SS = 6.9044$
$MS = 0.431525$

Note:
Try It
Exercise:

Problem:

Four sports teams took a random sample of players regarding their GPAs for the last year. The results are shown in [link].

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

GPAs FOR FOUR SPORTS TEAMS

Use a significance level of 5%, and determine if there is a difference in GPA among the teams.

Solution:

With a *p*-value of 0.9271, we decline to reject the null hypothesis. There is not sufficient evidence to conclude that there is a difference among the GPAs for the sports teams.

Example:
A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in [link].

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Exercise:

Problem:

Does it appear that the three media in which the bean plants were grown produce the same mean height? Test at a 3% level of significance.

Solution:

This time, we will perform the calculations that lead to the F' statistic. Notice that each group has the same number of plants, so we will use the formula $F' = \frac{n \cdot s_x{}^2}{s^2{}_{\text{pooled}}}$.

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	**Tara's Plants**	**Nick's Plants**
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Next, calculate the variance of the three group means (Calculate the variance of 24.2, 25.4, and 24.4). **Variance of the group means = 0.413 =** $s_x{}^2$

Then $MS_{between} = ns_x{}^2 = (5)(0.413)$ where $n = 5$ is the sample size (number of plants each child grew).

Calculate the mean of the three sample variances (Calculate the mean of 11.7, 18.3, and 16.3). **Mean of the sample variances = 15.433 =** s^2 ***pooled***

Then $MS_{within} = s^2{}_{pooled} = 15.433$.

The F statistic (or F ratio) is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{ns_x{}^2}{s^2{}_{pooled}} = \frac{(5)(0.413)}{15.433} = 0.134$

The *dfs* for the numerator = the number of groups – 1 = 3 – 1 = 2.

The *dfs* for the denominator = the total number of samples – the number of groups = 15 – 3 = 12

The distribution for the test is $F_{2,12}$ and the F statistic is $F = 0.134$

The p-value is $P(F > 0.134) = 0.8759$.

Decision: Since $\alpha = 0.03$ and the p-value = 0.8759, do not reject H_0. (Why?)

Conclusion: With a 3% level of significance, from the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

Note:
To calculate the p-value:
*Press `2nd DISTR`
*Arrow down to `Fcdf`(and press `ENTER`.
*Enter 0.134, `E99`, 2, 12)
*Press `ENTER`
The p-value is 0.8759.

Note:
Try It
Exercise:

Problem:

Another fourth grader also grew bean plants, but this time in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32. Do a one-way ANOVA test on the four groups. Are the heights of the bean plants different? Use the same method as shown in [link].

Solution:

- F = 0.9496
- p-value = 0.4402

From the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

Note:
Collaborative Exercise
From the class, create four groups of the same size as follows: men under 22, men at least 22, women under 22, women at least 22. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1% level of significance. Use one of the solution sheets in [link].

References

Data from a fourth grade classroom in 1994 in a private K – 12 school in San Jose, CA.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets: Data for Fruitfly Fecundity.* London: Chapman & Hall, 1994.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets.* London: Chapman & Hall, 1994, pg. 50.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. A Handbook of Small Datasets. London: Chapman & Hall, 1994, pg. 118.

“MLB Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/year/2012.

Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

Chapter Review

The graph of the F distribution is always positive and skewed right, though the shape can be mounded or exponential depending on the combination of numerator and denominator degrees of freedom. The F statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small F statistic will result, and the area under the F curve to the right will be large, representing a large p-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the

denominator, giving a large *F* statistic and a small area (small *p*-value) to the right of the statistic under the *F* curve.

When the data have unequal group sizes (unbalanced data), then techniques from [link] need to be used for hand calculations. In the case of balanced data (the groups are the same size) however, simplified calculations based on group means and variances may be used. In practice, of course, software is usually employed in the analysis. As in any analysis, graphs of various sorts should be used in conjunction with numerical techniques. Always look of your data!

Exercise:

Problem: An *F* statistic can have what values?

Exercise:

Problem:

What happens to the curves as the degrees of freedom for the numerator and the denominator get larger?

Solution:

The curves approximate the normal distribution.

Use the following information to answer the next seven exercise. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). The results are shown in [link].

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

Exercise:

Problem: What is the *df(num)*?

Exercise:

Problem: What is the *df(denom)*?

Solution:

ten

Exercise:

Problem: What are the Sum of Squares and Mean Squares Factors?

Exercise:

Problem: What are the Sum of Squares and Mean Squares Errors?

Solution:

SS = 237.33; *MS* = 23.73

Exercise:

Problem: What is the *F* statistic?

Exercise:

Problem: What is the *p*-value?

Solution:

0.1614

Exercise:

Problem: At the 5% significance level, is there a difference in the mean jump heights among the teams?

Use the following information to answer the next seven exercises. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in [link]

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

Exercise:

Problem: What is the *df(num)*?

Solution:

two

Exercise:

Problem: What is the *df(denom)*?

Exercise:

Problem: What are the $SS_{between}$ and $MS_{between}$?

Solution:

SS = 5,700.4;

MS = 2,850.2

Exercise:

Problem: What are the SS_{within} and MS_{within}?

Exercise:

Problem: What is the F Statistic?

Solution:

3.6101

Exercise:

Problem: What is the p-value?

Exercise:

Problem: At the 10% significance level, are the scores among the different groups different?

Solution:

Yes, there is enough evidence to show that the scores among the groups are statistically significant at the 10% level.

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	**South**	**West**	**Central**	**East**
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6

	Northeast	South	West	Central	East
	16.5	16.2	16.1	16.4	16.8
$x =$	________	________	________	________	________
$s^2 =$	________	________	________	________	________

Enter the data into your calculator or computer.

Exercise:

Problem: *p*-value = ______

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α.

Exercise:

Problem: $\alpha = 0.05$

a. Decision: ____________________________

b. Conclusion: ____________________________

Exercise:

Problem: $\alpha = 0.01$

a. Decision: ____________________________

b. Conclusion: ____________________________

Homework

Note:
DIRECTIONS
Use a solution sheet to conduct the following hypothesis tests. The solution sheet can be found in [link].

Exercise:

Problem:

Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 10%, test the hypothesis that the three formulas produce the same mean weight gain.

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Weights of Student Lab Rats

Solution:

a. H_0: $\mu_L = \mu_T = \mu_J$
b. H_a: at least any two of the means are different
c. $df(num) = 2$; $df(denom) = 12$
d. F distribution
e. 0.67
f. 0.5305
g. Check student's solution.
h. Decision: Do not reject null hypothesis; Conclusion: There is insufficient evidence to conclude that the means are different.

Exercise:

Problem:

A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in [link]. Using a 5% significance level, test the hypothesis that the three mean commuting mileages are the same.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6

working-class	professional (middle incomes)	professional (wealthy)
19.5	6.4	15.4
51.2	13.9	9.3

Use the following information to answer the next two exercises. [link] lists the number of pages in four different types of magazines.

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Exercise:

Problem:

Using a significance level of 5%, test the hypothesis that the four magazine types have the same mean length.

Exercise:

Problem:

Eliminate one magazine type that you now feel has a mean length different from the others. Redo the hypothesis test, testing that the remaining three means are statistically the same. Use a new solution sheet. Based on this test, are the mean lengths for the remaining three magazines statistically the same?

Solution:

a. H_a: $\mu_c = \mu_n = \mu_h$
b. At least any two of the magazines have different mean lengths.
c. $df(num) = 2$, $df(denom) = 12$
d. F distribtuion
e. $F = 15.28$
f. p-value = 0.0005
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the Null Hypothesis.
   iii. Reason for decision: p-value < alpha

iv. Conclusion: There is sufficient evidence to conclude that the mean lengths of the magazines are different.

Exercise:

Problem:

A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. Suppose that [link] shows the results of a study.

CNN	**FOX**	**Local**
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

Exercise:

Problem:

Are the means for the final exams the same for all statistics class delivery types? [link] shows the scores on final exams from several randomly selected classes that used the different delivery types.

Online	**Hybrid**	**Face-to-Face**
72	83	80
84	73	78
77	84	84
80	81	81
81		86

Online	Hybrid	Face-to-Face
		79
		82

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

Solution:

a. H_0: $\mu_o = \mu_h = \mu_f$
b. At least two of the means are different.
c. $df(n) = 2$, $df(d) = 13$
d. $F_{2,13}$
e. 0.64
f. 0.5437
g. Check student's solution.

h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: The mean scores of different class delivery are not different.

Exercise:

Problem:

Are the mean number of times a month a person eats out the same for whites, blacks, Hispanics and Asians? Suppose that [link] shows the results of a study.

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

Exercise:

Problem:

Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that [link] shows the results of a study.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

Solution:

a. H_0: $\mu_p = \mu_m = \mu_h$
b. At least any two of the means are different.
c. $df(n) = 2$, $df(d) = 12$
d. $F_{2,12}$
e. 3.13
f. 0.0807
g. Check student's solution.

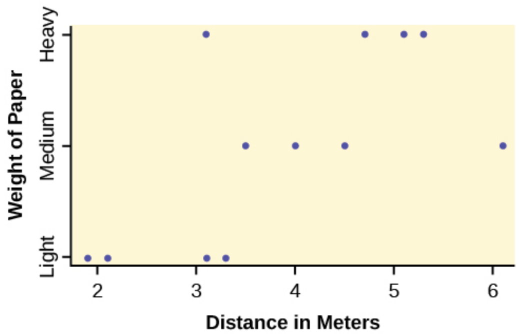
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: There is not sufficient evidence to conclude that the mean numbers of daily visitors are different.

Exercise:

Problem:

Sanjay made identical paper airplanes out of three different weights of paper, light, medium and heavy. He made four airplanes from each of the weights, and launched them himself across the room. Here are the distances (in meters) that his planes flew.

Paper Type/Trial	Trial 1	Trial 2	Trial 3	Trial 4
Heavy	5.1 meters	3.1 meters	4.7 meters	5.3 meters
Medium	4 meters	3.5 meters	4.5 meters	6.1 meters
Light	3.1 meters	3.3 meters	2.1 meters	1.9 meters



a. Take a look at the data in the graph. Look at the spread of data for each group (light, medium, heavy). Does it seem reasonable to assume a normal distribution with the same variance for each group? Yes or No.
b. Why is this a balanced design?
c. Calculate the sample mean and sample standard deviation for each group.
d. Does the weight of the paper have an effect on how far the plane will travel? Use a 1% level of significance. Complete the test using the method shown in the bean plant example in [link].

- variance of the group means ___________
- $MS_{between}$= ____________
- mean of the three sample variances ____________
- MS_{within} = ______________
- F statistic = _____________
- $df(num)$ = ___________, $df(denom)$ = ____________
- number of groups ________
- number of observations ________
- p-value = ___________ ($P(F >$ ________) = ___________)
- Graph the p-value.
- decision: _________________________
- conclusion: ____

Exercise:

Problem:

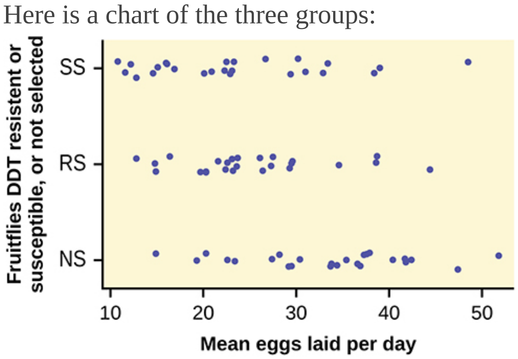
DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective, but persisted in the environment and over time became seen as harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.

An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). Finally there was a control line of non-selected or typical fruitflies (NS). Here are the data:

RS	SS	NS	RS	SS	NS
12.8	38.4	35.4	22.4	23.1	22.6
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
16.4	26.7	28.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
14.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1% level of significance, are the mean rates of egg selection for the three strains of fruitfly different? If so, in what way? Specifically, the researchers were interested in whether or not the selectively bred strains were different from the nonselected line, and whether the two selected lines were different from each other.

Here is a chart of the three groups:



Solution:

The data appear normally distributed from the chart and of similar spread. There do not appear to be any serious outliers, so we may proceed with our ANOVA calculations, to see if we have good evidence of a difference between the three groups.

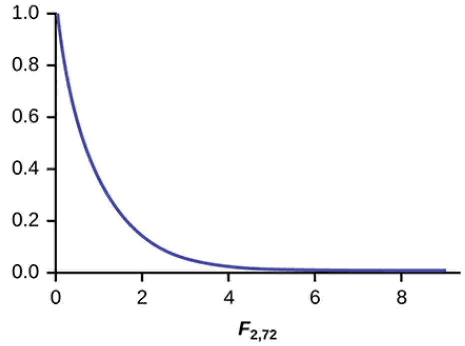
H_0: $\mu_1 = \mu_2 = \mu_3$;

H_a: $\mu_i \neq \mu_j$ some $i \neq j$.

Define μ_1, μ_2, μ_3, as the population mean number of eggs laid by the three groups of fruit flies.

F statistic = 8.6657;

p-value = 0.0004



Decision: Since the p-value is less than the level of significance of 0.01, we reject the null hypothesis.

Conclusion: We have good evidence that the average number of eggs laid during the first 14 days of life for these three strains of fruitflies are different.

Interestingly, if you perform a two sample t-test to compare the RS and NS groups they are significantly different (p = 0.0013). Similarly, SS and NS are significantly different (p = 0.0006). However, the two selected groups, RS and SS are *not* significantly different (p = 0.5176). Thus we appear to have good evidence that selection either for resistance or for susceptibility involves a reduced rate of egg production (for these specific strains) as compared to flies that were not selected for resistance or susceptibility to DDT. Here, genetic selection has apparently involved a loss of fecundity.

Exercise:

Problem:

The data shown is the recorded body temperatures of 130 subjects as estimated from available histograms.

Traditionally we are taught that the normal human body temperature is 98.6 F. This is not quite correct for everyone. Are the mean temperatures among the four groups different?

Calculate 95% confidence intervals for the mean body temperature in each group and comment about the confidence intervals.

FL	FH	ML	MH	FL	FH	ML	MH
96.4	96.8	96.3	96.9	98.4	98.6	98.1	98.6
96.7	97.7	96.7	97	98.7	98.6	98.1	98.6
97.2	97.8	97.1	97.1	98.7	98.6	98.2	98.7
97.2	97.9	97.2	97.1	98.7	98.7	98.2	98.8

FL	FH	ML	MH	FL	FH	ML	MH
97.4	98	97.3	97.4	98.7	98.7	98.2	98.8
97.6	98	97.4	97.5	98.8	98.8	98.2	98.8
97.7	98	97.4	97.6	98.8	98.8	98.3	98.9
97.8	98	97.4	97.7	98.8	98.8	98.4	99
97.8	98.1	97.5	97.8	98.8	98.9	98.4	99
97.9	98.3	97.6	97.9	99.2	99	98.5	99
97.9	98.3	97.6	98	99.3	99	98.5	99.2
98	98.3	97.8	98		99.1	98.6	99.5
98.2	98.4	97.8	98		99.1	98.6	
98.2	98.4	97.8	98.3		99.2	98.7	
98.2	98.4	97.9	98.4		99.4	99.1	
98.2	98.4	98	98.4		99.9	99.3	
98.2	98.5	98	98.6		100	99.4	
98.2	98.6	98	98.6		100.8		

Test of Two Variances

Another of the uses of the *F* distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

In order to perform a *F* test of two variances, it is important that the following are true:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

Unlike most other tests in this book, the *F* test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher *p*-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the *F* ratio:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]}$$

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F Ratio becomes

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}.$$

Note:
Note
The F ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on H_a and on which sample variance is larger.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to one. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than one. If s_1^2 and s_2^2 are far apart, then $F = \frac{(s_1)^2}{(s_2)^2}$ is a large number.

Therefore, if F is close to one, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than one, then the evidence is against the null hypothesis. **A test of two variances may be left, right, or two-tailed.**

Example:
Exercise:

Problem:

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

Solution:

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

$n_1 = n_2 = 30$.

H_0: $\sigma_1^2 = \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$

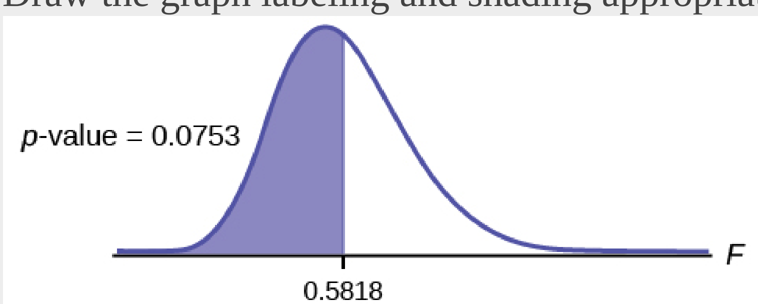
Calculate the test statistic: By the null hypothesis $\left(\sigma_1^2 = \sigma_2^2\right)$, the F statistic is:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818$$

Distribution for the test: $F_{29,29}$ where $n_1 – 1 = 29$ and $n_2 – 1 = 29$.

Graph: This test is left tailed.

Draw the graph labeling and shading appropriately.



Probability statement: p-value = $P(F < 0.5818) = 0.0753$

Compare α and the p-value: $\alpha = 0.10$ $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

Conclusion: With a 10% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Note:
Press `STAT` and arrow over to `TESTS`. Arrow down to `D:2-SampFTest`. Press `ENTER`. Arrow to `Stats` and press `ENTER`. For `Sx1`, `n1`, `Sx2`, and `n2`, enter $\sqrt{(52.3)}$, `30`, $\sqrt{(89.9)}$, and `30`. Press `ENTER` after each. Arrow to `σ1:` and `<σ2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. $F = 0.5818$ and p-value = 0.0753. Do the procedure again and try `Draw` instead of `Calculate`.

Note:
Try It
Exercise:

Problem:

The New York Choral Society divides male singers up into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the variance of the heights of singers in each of these two groups (Tenor1 and Bass2) are different?

Tenor1	**Bass2**	**Tenor 1**	**Bass 2**	**Tenor 1**	**Bass 2**
69	72	67	72	68	67
72	75	70	74	67	70
71	67	65	70	64	70
66	75	72	66		69
76	74	70	68		72
74	72	68	75		71
71	72	64	68		74
66	74	73	70		75
68	72	66	72		

Solution:

The histograms are not as normal as one might like. Plot them to verify. However, we proceed with the test in any case.

Subscripts: T1= tenor1 and B2 = bass 2

The standard deviations of the samples are s_{T1} = 3.3302 and s_{B2} = 2.7208.

The hypotheses are

$H_0 : \sigma_{T1}^2 = \sigma_{B2}^2$ and $H_0 : \sigma_{T1}^2 \neq \sigma_{B2}^2$ (two tailed test)

The F statistic is 1.4894 with 20 and 25 degrees of freedom.

The p-value is 0.3430. If we assume alpha is 0.05, then we cannot reject the null hypothesis.

We have no good evidence from the data that the heights of Tenor1 and Bass2 singers have different variances (despite there being a significant difference in mean heights of about 2.5 inches.)

References

“MLB Vs. Division Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true.

Chapter Review

The F test for the equality of two variances rests heavily on the assumption of normal distributions. The test is unreliable if this assumption is not met. If both distributions are normal, then the ratio of the two sample variances is distributed as an F statistic, with numerator and denominator degrees of freedom that are one less than the samples sizes of the corresponding two

groups. A **test of two variances** hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the *F* distribution with two different degrees of freedom.

Assumptions:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

Formula Review

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

If $\sigma_1 = \sigma_2$, then $F = \frac{{s_1}^2}{{s_2}^2}$

Use the following information to answer the next two exercises. There are two assumptions that must be true in order to perform an *F* test of two variances.

Exercise:

Problem: Name one assumption that must be true.

Solution:

The populations from which the two samples are drawn are normally distributed.

Exercise:

Problem: What is the other assumption that must be true?

Use the following information to answer the next five exercises. Two coworkers commute from the same building. They are interested in whether or not there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker thinks that he is more consistent with his commute times. Test the claim at the 10% level. Assume that commute times are normally distributed.

Exercise:

Problem: State the null and alternative hypotheses.

Solution:

H_0: $\sigma_1 = \sigma_2$

H_a: $\sigma_1 < \sigma_2$

or

H_0: $\sigma_1^2 = \sigma_2^2$

H_a: $\sigma_1^2 < \sigma_2^2$

Exercise:

Problem: What is s_1 in this problem?

Exercise:

Problem: What is s_2 in this problem?

Solution:

4.11

Exercise:

Problem: What is n?

Exercise:

Problem: What is the F statistic?

Solution:

0.7159

Exercise:

Problem: What is the p-value?

Exercise:

Problem: Is the claim accurate?

Solution:

No, at the 10% level of significance, we do not reject the null hypothesis and state that the data do not show that the variation in drive times for the first worker is less than the variation in drive times for the second worker.

Use the following information to answer the next four exercises. Two students are interested in whether or not there is variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. The second student thinks his scores are more consistent.

Exercise:

Problem: State the null and alternative hypotheses.

Exercise:

Problem: What is the *F* Statistic?

Solution:

2.8674

Exercise:

Problem: What is the *p*-value?

Exercise:

Problem:

At the 5% significance level, do we reject the null hypothesis?

Solution:

Reject the null hypothesis. There is enough evidence to say that the variance of the grades for the first student is higher than the variance in the grades for the second student.

Use the following information to answer the next three exercises. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8 and the second cyclist has a variance of 32.1. The cyclists want to see if their variances are the same or different. Assume that commute times are normally distributed.

Exercise:

Problem: State the null and alternative hypotheses.

Exercise:

Problem: What is the *F* Statistic?

Solution:

0.7414

Exercise:

Problem:

At the 5% significance level, what can we say about the cyclists' variances?

Homework

Exercise:

Problem:

Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0

Linda's rats	Tuan's rats	Javier's rats
38.2	44.2	48.6

Determine whether or not the variance in weight gain is statistically the same among Javier’s and Linda’s rats. Test at a significance level of 10%.

Solution:

a. H_0: $\sigma_1^2 = \sigma_2^2$
b. H_a: $\sigma_1^2 \neq \sigma_1^2$
c. *df*(*num*) = 4; *df*(*denom*) = 4
d. $F_{4, 4}$
e. 3.00
f. 2(0.1563) = 0.3126. Using the TI-83+/84+ function 2-SampFtest, you get the test statistic as 2.9986 and *p*-value directly as 0.3127. If you input the lists in a different order, you get a test statistic of 0.3335 but the *p*-value is the same because this is a two-tailed test.
g. Check student't solution.
h. Decision: Do not reject the null hypothesis; Conclusion: There is insufficient evidence to conclude that the variances are different.

Exercise:

Problem:

A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are as follows.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Determine whether or not the variance in mileage driven is statistically the same among the working class and professional (middle income) groups. Use a 5% significance level.

Exercise:

Problem:

Which two magazine types do you think have the same variance in length?

Exercise:

Problem:

Which two magazine types do you think have different variances in length?

Solution:

The answers may vary. Sample answer: Home decorating magazines and news magazines have different variances.

Exercise:

Problem:

Is the variance for the amount of money, in dollars, that shoppers spend on Saturdays at the mall the same as the variance for the amount of money that shoppers spend on Sundays at the mall? Suppose that the [link] shows the results of a study.

Saturday	Sunday	Saturday	Sunday
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

Exercise:

Problem:

Are the variances for incomes on the East Coast and the West Coast the same? Suppose that [link] shows the results of a study. Income is shown in thousands of dollars. Assume that both distributions are normal. Use a level of significance of 0.05.

East	West
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

Solution:

a. H_0: $= \sigma_1^2 = \sigma_2^2$
b. H_a: $\sigma_1^2 \neq \sigma_1^2$
c. $df(n) = 7$, $df(d) = 6$

d. $F_{7,6}$
e. 0.8117
f. 0.7825
g. Check student’s solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: p-value > alpha
  iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

Exercise:

Problem:

Thirty men in college were taught a method of finger tapping. They were randomly assigned to three groups of ten, with each receiving one of three doses of caffeine: 0 mg, 100 mg, 200 mg. This is approximately the amount in no, one, or two cups of coffee. Two hours after ingesting the caffeine, the men had the rate of finger tapping per minute recorded. The experiment was double blind, so neither the recorders nor the students knew which group they were in. Does caffeine affect the rate of tapping, and if so how?

Here are the data:

0 mg	**100 mg**	**200 mg**	**0 mg**	**100 mg**	**200 mg**
242	248	246	245	246	248
244	245	250	248	247	252

0 mg	100 mg	200 mg	0 mg	100 mg	200 mg
247	248	248	248	250	250
242	247	246	244	246	248
246	243	245	242	244	250

Exercise:

Problem:

King Manuel I, Komnenus ruled the Byzantine Empire from Constantinople (Istanbul) during the years 1145 to 1180 A.D. The empire was very powerful during his reign, but declined significantly afterwards. Coins minted during his era were found in Cyprus, an island in the eastern Mediterranean Sea. Nine coins were from his first coinage, seven from the second, four from the third, and seven from a fourth. These spanned most of his reign. We have data on the silver content of the coins:

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

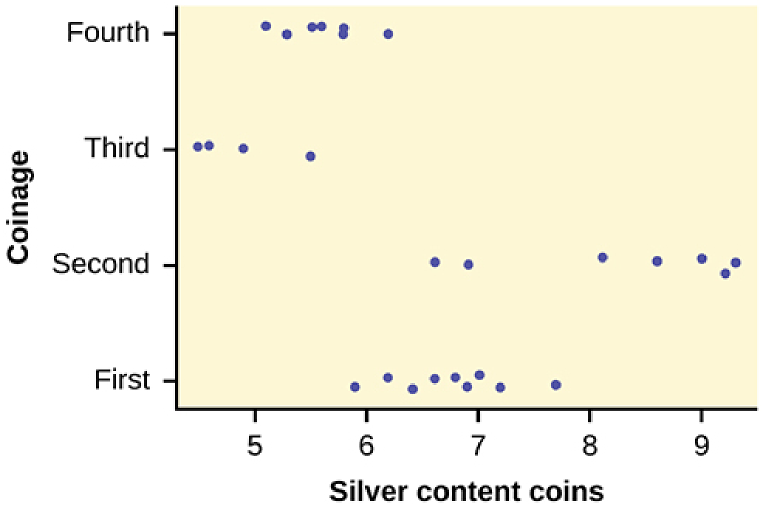
Did the silver content of the coins change over the course of Manuel's reign?

Here are the means and variances of each coinage. The data are unbalanced.

	First	Second	Third	Fourth
Mean	6.7444	8.2429	4.875	5.6143
Variance	0.2953	1.2095	0.2025	0.1314

Solution:

Here is a strip chart of the silver content of the coins:



While there are differences in spread, it is not unreasonable to use ANOVA techniques. Here is the completed ANOVA table:

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	37.748	4 – 1 = 3	12.5825	26.272
Error (Within)	11.015	27 – 4 = 23	0.4789	
Total	48.763	27 – 1 = 26		

$P(F > 26.272) = 0$;

Reject the null hypothesis for any alpha. There is sufficient evidence to conclude that the mean silver content among the four coinages are different. From the strip chart, it appears that the first and second coinages had higher silver contents than the third and fourth.

Exercise:

Problem:

The American League and the National League of Major League Baseball are each divided into three divisions: East, Central, and West. Many years, fans talk about some divisions being stronger (having better teams) than other divisions. This may have consequences for the postseason. For instance, in 2012 Tampa Bay won 90 games and did not play in the postseason, while Detroit won only 88 and did play in the postseason. This may have been an oddity, but is there good evidence that in the 2012 season, the American League divisions were significantly different in overall records? Use the following data to test whether the mean number of wins per team in the three American League divisions were the same or not. Note that the data are not balanced, as two divisions had five teams, while one had only four.

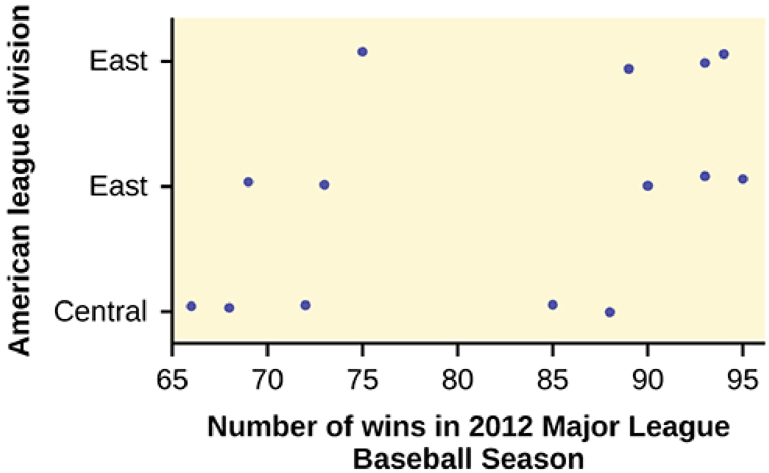
Division	**Team**	**Wins**
East	NY Yankees	95
East	Baltimore	93
East	Tampa Bay	90
East	Toronto	73
East	Boston	69

Division	Team	Wins
Central	Detroit	88
Central	Chicago Sox	85
Central	Kansas City	72
Central	Cleveland	68
Central	Minnesota	66

Division	Team	Wins
West	Oakland	94
West	Texas	93
West	LA Angels	89
West	Seattle	75

Solution:

Here is a stripchart of the number of wins for the 14 teams in the AL for the 2012 season.



While the spread seems similar, there may be some question about the normality of the data, given the wide gaps in the middle near the 0.500 mark of 82 games (teams play 162 games each season in MLB). However, one-way ANOVA is robust.

Here is the ANOVA table for the data:

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	344.16	3 – 1 = 2	172.08	
Error (Within)	1,219.55	14 – 3 = 11	110.87	1.5521
Total	1,563.71	14 – 1 = 13		

$P(F > 1.5521) = 0.2548$
Since the p-value is so large, there is not good evidence against the null hypothesis of equal means. We decline to reject the null hypothesis. Thus, for 2012, there is not any have any good evidence of a significant difference in mean number of wins between the divisions of the American League.

Lab: One-Way ANOVA

Note:
One-Way ANOVA
Class Time:
Names:
Student Learning Outcome

- The student will conduct a simple one-way ANOVA test involving three variables.

Collect the Data

1. Record the price per pound of eight fruits, eight vegetables, and eight breads in your local supermarket.

Fruits	Vegetables	Breads

2. Explain how you could try to collect the data randomly.

Analyze the Data and Conduct a Hypothesis Test

1. State the null hypothesis and the alternative hypothesis.

2. Compute the following:

   a. Fruit:

      i. x = _______
      ii. s_x = _______
      iii. n = _______

   b. Vegetables:

      i. x = _______
      ii. s_x = _______
      iii. n = _______

   c. Bread:

      i. x = _______
      ii. s_x = _______
      iii. n = _______

3. Find the following:

   a. $df(num)$ = _______
   b. $df(denom)$ = _______

4. State the approximate distribution for the test.
5. Test statistic: F = _______
6. Sketch a graph of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.
7. *p*-value = _______
8. Test at $\alpha = 0.05$. State your decision and conclusion.

9. a. Decision: Why did you make this decision?
   b. Conclusion (write a complete sentence).
   c. Based on the results of your study, is there a need to investigate any of the food groups' prices? Why or why not?

Review Exercises (Ch 3-13)

These review exercises are designed to provide extra practice on concepts learned before a particular chapter. For example, the review exercises for Chapter 3, cover material learned in chapters 1 and 2.

Chapter 3

Use the following information to answer the next six exercises: In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks.

1. The “average increase” for all NASDAQ stocks is the:

a. population
b. statistic
c. parameter
d. sample
e. variable

2. All of the NASDAQ stocks are the:

a. population
b. statistics
c. parameter
d. sample
e. variable

3. Nine percent is the:

a. population
b. statistics
c. parameter
d. sample
e. variable

4. The 100 NASDAQ stocks in the survey are the:

a. population
b. statistic
c. parameter
d. sample
e. variable

5. The percent increase for one stock in the survey is the:

a. population
b. statistic
c. parameter
d. sample
e. variable

6. Would the data collected by qualitative, quantitative discrete, or quantitative continuous?

Use the following information to answer the next two exercises: Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Weight Gain	**Frequency**
–2	3
–1	5
0	2
1	4
4	13
6	2

Weight Gain	Frequency
11	1

7. Calculate the following values:

a. the average weight gain for the two weeks
b. the standard deviation
c. the first, second, and third quartiles

8. Construct a histogram and box plot of the data.

Chapter 4

Use the following information to answer the next two exercises: A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2,000 per month. Of those respondents who charge more than $2,000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

9. What is the probability that a randomly selected respondent will spend more than $2,000 AND use a credit card that gives them a mile of air travel for every dollar they charge?

a. (0.30)(0.35)
b. (0.80)(0.35)
c. (0.80)(0.30)
d. (0.80)

10. Are using a credit card that gives a mile of air travel for each dollar spent AND charging more than $2,000 per month independent events?

a. Yes
b. No, and they are not mutually exclusive either.
c. No, but they are mutually exclusive.
d. Not enough information given to determine the answer

11. A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Sixty-eight questionnaires are returned. What is the population in this study?

a. all employed adult women
b. all the members of a local business and professional women's club
c. the 100 women who received the questionnaire
d. all employed women with children

Use the following information to answer the next two exercises: The next two questions refer to the following: An article from The San Jose Mercury News was concerned with the racial mix of the 1500 students at Prospect High School in Saratoga, CA. The table summarizes the results. (Male and female values are approximate.) Suppose one Prospect High School student is randomly selected.

Gender/Ethnic group	**White**	**Asian**	**Hispanic**	**Black**	**American Indian**
Male	400	468	115	35	16
Female	440	132	140	40	14

12. Find the probability that a student is Asian or Male.

13. Find the probability that a student is Black given that the student is female.

14. A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs.
- Median = 4.5 lbs.

- Mode = 4 lbs.
- Standard deviation = 3.8 lbs.
- First quartile = 2 lbs.
- Third quartile = 8.5 lbs.

The correct statement is:

a. One fourth of the members lost exactly two pounds.
b. The middle fifty percent of the members lost from two to 8.5 lbs.
c. Most people lost 3.5 to 4.5 lbs.
d. All of the choices above are correct.

15. What does it mean when a data set has a standard deviation equal to zero?

a. All values of the data appear with the same frequency.
b. The mean of the data is also zero.
c. All of the data have the same value.
d. There are no data to begin with.

16. The statement that describe the illustration is:

a. the mean is equal to the median.
b. There is no first quartile.
c. The lowest data value is the median.
d. The median equals $\frac{Q_1+Q_3}{2}$.

17. According to a recent article in the *San Jose Mercury News* the average number of babies born with significant hearing loss (deafness) is approximately 2 per 1000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby

nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

18. A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn’t matter. I expect to break even.

Use the following information to answer the next four exercises: Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he/she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

19. Define the random variable and list its possible values.

20. State the distribution of X.

21. Find the probability that at least four of the 25 patients actually have the flu.

22. On average, for every 25 patients calling in, how many do you expect to have the flu?

Use the following information to answer the next two exercises: Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

23. What kind of data was collected?

a. qualitative
b. quantitative continuous
c. quantitative discrete

24. What is the population under study?

Chapter 5

Use the following information to answer the next seven exercises: A recent study of mothers of junior high school children in Santa Clara County reported that 76% of the mothers are employed in paid positions. Of those mothers who are employed, 64% work full-time (over 35 hours per week), and 36% work part-time. However, out of all of the mothers in the population, 49% work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County. Let E = employed and F = full-time employment.

25.

a. Find the percent of all mothers in the population that are NOT employed.
b. Find the percent of mothers in the population that are employed part-time.

26. The "type of employment" is considered to be what type of data?

27. Find the probability that a randomly selected mother works part-time given that she is employed.

28. Find the probability that a randomly selected person from the population will be employed or work full-time.

29. Being employed and working part-time:

a. mutually exclusive events? Why or why not?
b. independent events? Why or why not?

Use the following additional information to answer the next two exercises: We randomly pick ten mothers from the above population. We are interested in the

number of the mothers that are employed. Let X = number of mothers that are employed.

30. State the distribution for X.

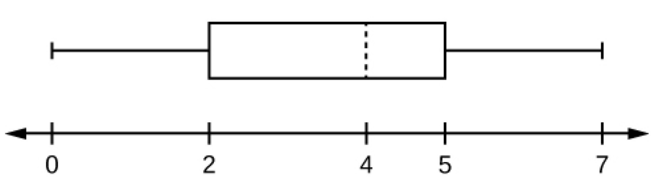
31. Find the probability that at least six are employed.

32. We expect the statistics discussion board to have, on average, 14 questions posted to it per week. We are interested in the number of questions posted to it per day.

a. Define X.
b. What are the values that the random variable may take on?
c. State the distribution for X.
d. Find the probability that from ten to 14 (inclusive) questions are posted to the listserv on a randomly picked day.

33. A person invests $1,000 into stock of a company that hopes to go public in one year. The probability that the person will lose all his money after one year (i.e. his stock will be worthless) is 35%. The probability that the person's stock will still have a value of $1,000 after one year (i.e. no profit and no loss) is 60%. The probability that the person's stock will increase in value by $10,000 after one year (i.e. will be worth $11,000) is 5%. Find the expected profit after one year.

34. Rachel's piano cost $3,000. The average cost for a piano is $4,000 with a standard deviation of $2,500. Becca's guitar cost $550. The average cost for a guitar is $500 with a standard deviation of $200. Matt's drums cost $600. The average cost for drums is $700 with a standard deviation of $100. Whose cost was lowest when compared to his or her own instrument?

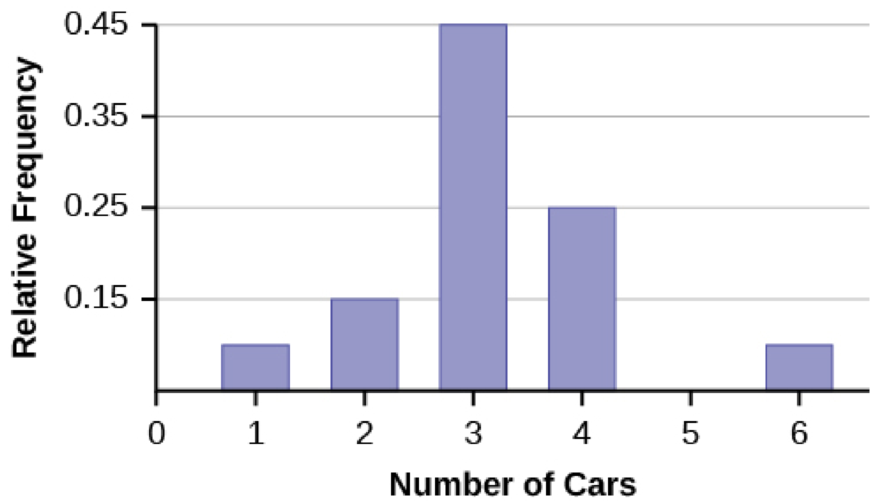


35. Explain why each statement is either true or false given the box plot in [link].

a. Twenty-five percent of the data re at most five.
b. There is the same amount of data from 4–5 as there is from 5–7.
c. There are no data values of three.

d. Fifty percent of the data are four.

Using the following information to answer the next two exercises: 64 faculty members were asked the number of cars they owned (including spouse and children's cars). The results are given in the following graph:



36. Find the approximate number of responses that were three.

37. Find the first, second and third quartiles. Use them to construct a box plot of the data.

Use the following information to answer the next three exercises: [link] shows data gathered from 15 girls on the Snow Leopard soccer team when they were asked how they liked to wear their hair. Supposed one girl from the team is randomly selected.

Hair Style/Hair Color	**Blond**	**Brown**	**Black**
Ponytail	3	2	5
Plain	2	2	1

38. Find the probability that the girl has black hair GIVEN that she wears a ponytail.

39. Find the probability that the girl wears her hair plain OR has brown hair.

40. Find the probability that the girl has blond hair AND that she wears her hair plain.

Chapter 6

Use the following information to answer the next two exercises: $X \sim U(3, 13)$

41. Explain which of the following are false and which are true.

a. $f(x) = \frac{1}{10}$, $3 \le x \le 13$
b. There is no mode
c. The median is less than the mean.
d. $P(x > 10) = P(x \le 6)$

42. Calculate:

a. the mean.
b. the median.
c. the 65th percentile.

43. Which of the following is true for the box plot in [link]?

a. Twenty-five percent of the data are at most five.
b. There is about the same amount of data from 4–5 as there is from 5–7.
c. There are no data values of three.
d. Fifty percent of the data are four.

44. If $P(G|H) = P(G)$, then which of the following is correct?

a. G and H are mutually exclusive events.
b. $P(G) = P(H)$
c. Knowing that H has occurred will affect the chance that G will happen.
d. G and H are independent events.

45. If $P(J) = 0.3$, $P(K) = 0.63$, and J and K are independent events, then explain which are correct and which are incorrect.

a. $P(J \text{ AND } K) = 0$
b. $P(J \text{ OR } K) = 0.9$
c. $P(J \text{ OR } K) = 0.72$
d. $P(J) \neq P(J|K)$

46. On average, five students from each high school class get full scholarships to four-year colleges. Assume that most high school classes have about 500 students. X = the number of students from a high school class that get full scholarships to four-year schools. Which of the following is the distribution of X?

a. $P(5)$
b. $B(500, 5)$
c. $Exp\left(\frac{1}{5}\right)$
d. $N\left(5, \frac{(0.01)(0.99)}{500}\right)$

Chapter 7

Use the following information to answer the next three exercises: Richard’s Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

47. X ~ __________

a. $U(0, 4)$
b. $U(10, 20)$

c. *Exp*(2)
d. *N*(2, 1)

48. The average wait time is:

a. 1 hour.
b. 2 hours.
c. 2.5 hours.
d. 4 hours.

49. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least 1.5 more hours is:

a. $\frac{1}{4}$
b. $\frac{1}{2}$
c. $\frac{3}{4}$
d. $\frac{3}{8}$

50. Given: $X \sim Exp\left(\frac{1}{3}\right)$

a. Find $P(x > 1)$.
b. Calculate the minimum value for the upper quartile.
c. Find $P\left(x = \frac{1}{3}\right)$

51.

- 40% of full-time students took 4 years to graduate
- 30% of full-time students took 5 years to graduate
- 20% of full-time students took 6 years to graduate
- 10% of full-time students took 7 years to graduate

The expected time for full-time students to graduate is:

a. 4 years

b. 4.5 years
c. 5 years
d. 5.5 years

52. Which of the following distributions is described by the following example? Many people can run a short distance of under two miles, but as the distance increases, fewer people can run that far.

a. binomial
b. uniform
c. exponential
d. normal

53. The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his or her teeth less than $\frac{3}{4}$ minutes.

a. 0.5
b. $\frac{3}{4}$
c. 0.43
d. 0.63

54. Which distribution accurately describes the following situation?
The chance that a teenage boy regularly gives his mother a kiss goodnight is about 20%. Fourteen teenage boys are randomly surveyed. Let X = the number of teenage boys that regularly give their mother a kiss goodnight.

a. $B(14,0.20)$
b. $P(2.8)$
c. $N(2.8,2.24)$
d. $Exp\left(\frac{1}{0.20}\right)$

55. A 2008 report on technology use states that approximately 20% of U.S. households have never sent an e-mail. Suppose that we select a random sample of

fourteen U.S. households. Let X = the number of households in a 2008 sample of 14 households that have never sent an email

a. $B(14,0.20)$
b. $P(2.8)$
c. $N(2.8,2.24)$
d. $Exp\left(\frac{1}{0.20}\right)$

Chapter 8

Use the following information to answer the next three exercises: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to three pounds. Assume that the distribution for the weight loss is normal.

56. To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be:

a. number of people who lost weight on the special weight loss diet.
b. the number of people who were on the diet.
c. the mean amount of weight lost by 15 people on the special weight loss diet.
d. the total amount of weight lost by 15 people on the special weight loss diet.

57. Find the probability asked for in Question 56.

58. Find the 90th percentile for the mean amount of weight lost by 15 people.

Using the following information to answer the next three exercises: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time (hours) it takes for the first accident to occur.

59. What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

a. cannot be determined from the information given
b. $\frac{1}{6}$

c. $\frac{1}{2}$
d. $\frac{1}{3}$

60. The 20^{th} percentile occurs after how many hours?

a. 0.20
b. 0.60
c. 0.50
d. 1

61. Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let C = the total cumulative time. Then C follows which distribution?

a. $U(0,3)$
b. $Exp(13)$
c. $N(60, 5.477)$
d. $N(1.5, 0.01875)$

62. Using the information in Question 61, find the probability that the total time for all first accidents to occur is more than 43 hours.

Use the following information to answer the next two exercises: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from one to 15 days.

63. How long must a parent expect to wait for his children to clean their rooms?

a. eight days
b. three days
c. 14 days
d. six days

64. What is the probability that a parent will wait more than six days given that the parent has already waited more than three days?

a. 0.5174
b. 0.0174
c. 0.7500
d. 0.2143

Use the following information to answer the next five exercises: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

65. Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

a. 80%
b. 20%
c. 30%
d. cannot be determined

66. Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

a. 50%
b. 35%
c. 27.5%
d. 75%

67. Are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

a. yes
b. no
c. cannot be determined

68. The interest rate charged on the financial aid is ________ data.

a. quantitative discrete
b. quantitative continuous
c. qualitative discrete
d. qualitative

69. The following information is about the students who receive financial aid at the local community college.

- 1st quartile = $250
- 2nd quartile = $700
- 3rd quartile = $1200

These amounts are for the school year. If a sample of 200 students is taken, how many are expected to receive $250 or more?

a. 50
b. 250
c. 150
d. cannot be determined

Use the following information to answer the next two exercises: $P(A) = 0.2$, $P(B) = 0.3$; A and B are independent events.

70. $P(A \text{ AND } B)$ = _______

a. 0.5
b. 0.6
c. 0
d. 0.06

71. $P(A \text{ OR } B)$ = ________

a. 0.56
b. 0.5
c. 0.44
d. 1

72. If H and D are mutually exclusive events, $P(H) = 0.25$, $P(D) = 0.15$, then $P(H|D)$.

a. 1
b. 0
c. 0.40
d. 0.0375

Chapter 9

73. Rebecca and Matt are 14 year old twins. Matt’s height is two standard deviations below the mean for 14 year old boys’ height. Rebecca’s height is 0.10 standard deviations above the mean for 14 year old girls’ height. Interpret this.

a. Matt is 2.1 inches shorter than Rebecca.
b. Rebecca is very tall compared to other 14 year old girls.
c. Rebecca is taller than Matt.
d. Matt is shorter than the average 14 year old boy.

74. Construct a histogram of the IPO data (see [link]).

Use the following information to answer the next three exercises: Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. Let X = the number of estimates.

x	Relative Frequency	Cumulative Relative Frequency
1	0.3	
2	0.2	
4	0.4	

x	Relative Frequency	Cumulative Relative Frequency
5	0.1	

75. Complete the cumulative frequency column.

76. Calculate the sample mean (a), the sample standard deviation (b) and the percent of the estimates that fall at or below four (c).

77. Calculate the median, M, the first quartile, Q_1, the third quartile, Q_3. Then construct a box plot of the data.

78. The middle 50% of the data are between _____ and _____.

Use the following information to answer the next three exercises: Seventy 5th and 6th graders were asked their favorite dinner.

	Pizza	Hamburgers	Spaghetti	Fried shrimp
5th grader	15	6	9	0
6th grader	15	7	10	8

79. Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

a. $\frac{32}{70}$
b. $\frac{8}{32}$
c. $\frac{8}{8}$
d. $\frac{8}{70}$

80. Find the probability that a child does not prefer pizza.

a. $\frac{30}{70}$
b. $\frac{30}{40}$
c. $\frac{40}{70}$
d. 1

81. Find the probability a child is in the 5^{th} grade given that the child prefers spaghetti.

a. $\frac{9}{19}$
b. $\frac{9}{70}$
c. $\frac{9}{30}$
d. $\frac{19}{70}$

82. A sample of convenience is a random sample.

a. true
b. false

83. A statistic is a number that is a property of the population.

a. true
b. false

84. You should always throw out any data that are outliers.

a. true
b. false

85. Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on average. Of interest is the number of pies she bakes each day.

a. Define the random variable *X*.

b. State the distribution for *X*.
c. Find the probability that Lee bakes more than 25 pies in any given day.

86. Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, 5. Assume that the underlying distribution is normal. Calculate a 95% confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

87. Given: uniform, exponential, normal distributions. Match each to a statement below.

a. mean = median ≠ mode
b. mean > median > mode
c. mean = median = mode

Chapter 10

Use the following information to answer the next three exercises: In a survey at Kirkwood Ski Resort the following information was recorded:

	0–10	**11–20**	**21–40**	**40+**
Ski	10	12	30	8
Snowboard	6	17	12	5

Suppose that one person from [link] was randomly selected.

88. Find the probability that the person was a skier or was age 11–20.

89. Find the probability that the person was a snowboarder given he or she was age 21–40.

90. Explain which of the following are true and which are false.

a. Sport and age are independent events.
b. Ski and age 11–20 are mutually exclusive events.
c. *P*(Ski AND age 21–40) < *P*(Ski|age 21–40)
d. *P*(Snowboard OR age 0–10) < *P*(Snowboard|age 0–10)

91. The average length of time a person with a broken leg wears a cast is approximately six weeks. The standard deviation is about three weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the thirty people.

92. The distribution for X is uniform. What can we say for certain about the distribution for X when $n = 1$?

a. The distribution for X is still uniform with the same mean and standard deviation as the distribution for X.
b. The distribution for X is normal with the different mean and a different standard deviation as the distribution for X.
c. The distribution for X is normal with the same mean but a larger standard deviation than the distribution for X.
d. The distribution for X is normal with the same mean but a smaller standard deviation than the distribution for X.

93. The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

a. distribution for $\sum X$ is still uniform with the same mean and standard deviation as the distribution for X.
b. The distribution for $\sum X$ is normal with the same mean but a larger standard deviation as the distribution for X.
c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.
d. The distribution for $\sum X$ is normal with the same mean but a smaller standard deviation than the distribution for X.

Use the following information to answer the next three exercises: A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

94. State the approximate distribution for X, the distribution for the average lengths of baby carrots in 16 five-pound bags. X ~ ______

95. Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

96. Find the probability that x is between two and 2.25 inches.

Use the following information to answer the next three exercises: At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of five minutes and a standard deviation of two minutes.

97. Find the 90th percentile of waiting time in minutes.

98. Find the median waiting time for one student.

99. Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Chapter 11

Use the following information to answer the next four exercises: Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of seven years and a standard deviation of two years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

100. Sixty percent of individuals keep their cars **at most** how many years?

101. Suppose that we randomly survey one person. Find the probability that person keeps his or her car **less than** 2.5 years.

102. If we are to pick individuals ten at a time, find the distribution for the **mean** car length ownership.

103. If we are to pick ten individuals, find the probability that the **sum** of their ownership time is more than 55 years.

104. For which distribution is the median not equal to the mean?

a. Uniform
b. Exponential
c. Normal
d. Student *t*

105. Compare the standard normal distribution to the Student's *t*-distribution, centered at zero. Explain which of the following are true and which are false.

a. As the number surveyed increases, the area to the left of –1 for the Student's *t*-distribution approaches the area for the standard normal distribution.
b. As the degrees of freedom decrease, the graph of the Student's *t*-distribution looks more like the graph of the standard normal distribution.
c. If the number surveyed is 15, the normal distribution should never be used.

Use the following information to answer the next five exercises: We are interested in the checking account balance of twenty-year-old college students. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of $640 and a sample standard deviation of $150. Let X = checking account balance of an individual twenty year old college student.

106. Explain why we cannot determine the distribution of X.

107. If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of twenty-year-old college students, what distribution would you use?

108. Find the 95% confidence interval for the true mean checking account balance of a twenty-year-old college student.

109. What type of data is the balance of the checking account considered to be?

110. What type of data is the number of twenty-year-olds considered to be?

111. On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

a. Define the random variable X.

b. State the distribution for *X*.
c. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Use the following information to answer the next two exercises: The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30. Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

112. Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

113. Is it likely that exactly eight of the 15 plays would pay back money? Justify your answer numerically.

114. A game is played with the following rules:

- it costs $10 to enter.
- a fair coin is tossed four times.
- if you do not get four heads or four tails, you lose your $10.
- if you get four heads or four tails, you get back your $10, plus $30 more.

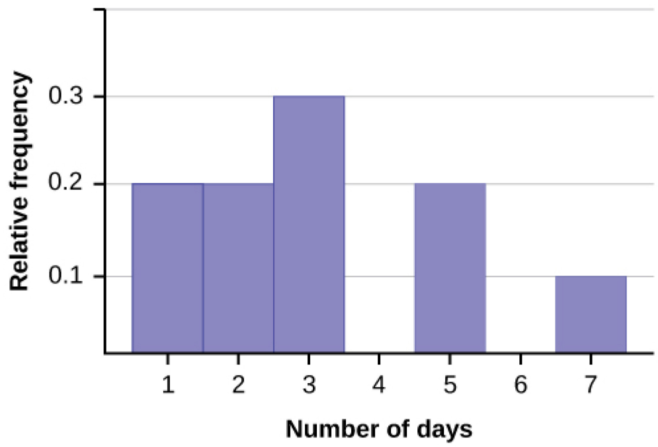
Over the long run of playing this game, what are your expected earnings?

115.

- The mean grade on a math exam in Rachel's class was 74, with a standard deviation of five. Rachel earned an 80.
- The mean grade on a math exam in Becca's class was 47, with a standard deviation of two. Becca earned a 51.
- The mean grade on a math exam in Matt's class was 70, with a standard deviation of eight. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Use the following information to answer the next two exercises: A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:



116. Find the number of responses that were five.

117. Find the mean, standard deviation, the median, the first quartile, the third quartile and the *IQR*.

118. Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home. Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

a. State an appropriate null hypothesis.
b. State an appropriate alternative hypothesis.
c. Define the random variable, P'.
d. Calculate the test statistic.
e. Calculate the *p*-value.
f. At the 5% level of decision, what is your decision about the null hypothesis?
g. What is the Type I error?
h. What is the Type II error?

119. Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

120. It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. Fifteen Californians wear jeans

to work and six non-Californians wear jeans to work.
Let *C* = Californian professional; *NC* = non-Californian professional

a. State appropriate null and alternate hypotheses.
b. Define the random variable.
c. Calculate the test statistic and *p*-value.
d. At the 5% significance level, what is your decision?
e. What is the Type I error?
f. What is the Type II error?

Use the following information to answer the next two exercises: A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

121. This is a test of (pick the best answer):

a. large samples, independent means
b. small samples, independent means
c. dependent means

122. State the distribution to use for the test.

Chapter 12

Use the following information to answer the next two exercises: A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12–19. Six percent of the girls surveyed said they have been pregnant. We are interested in the true proportion of U.S. girls, age 12–19, who have been pregnant.

123. Find the 95% confidence interval for the true proportion of U.S. girls, age 12–19, who have been pregnant.

124. The report also stated that the results of the survey are accurate to within ±3.7% at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What is the minimum number that should be surveyed?

125. Given: $X \sim Exp\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

Use the following information to answer the next three exercises: The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the mean amount of money a customer spends in one trip to the supermarket is $72.

126. Find the probability that one customer spends less than $72 in one trip to the supermarket?

127. Suppose five customers pool their money. How much money altogether would you expect the five customers to spend in one trip to the supermarket (in dollars)?

128. State the distribution to use if you want to find the probability that the **mean** amount spent by five customers in one trip to the supermarket is less than $60.

Chapter 13

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

129. What is the probability of both a drought **and** water rationing occurring?

130. Out of the years with water rationing, find the probability that there is a drought.

Use the following information to answer the next three exercises:

	Apple	**Pumpkin**	**Pecan**
Female	40	10	30
Male	20	30	10

131. Suppose that one individual is randomly chosen. Find the probability that the person’s favorite pie is apple **or** the person is male.

132. Suppose that one male is randomly chosen. Find the probability his favorite pie is pecan.

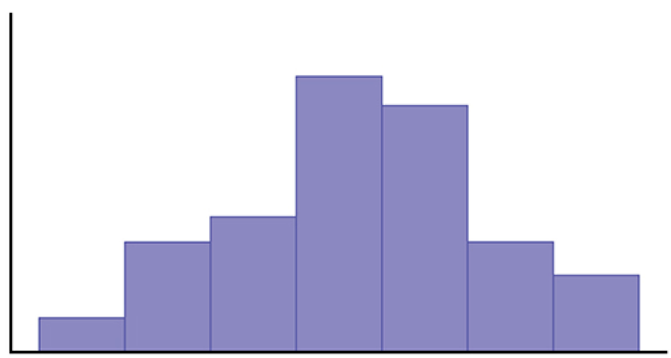
133. Conduct a hypothesis test to determine if favorite pie type and gender are independent.

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60.

134. We randomly survey 14 people. On average, how many people do we expect to watch the news at least once per week?

135. We randomly survey 14 people. Of interest is the number that watch the news at least once per week. State the distribution of X. $X \sim$ ______

136. The following histogram is most likely to be a result of sampling from which distribution?



a. Chi-Square
b. Geometric
c. Uniform
d. Binomial

137. The ages of De Anza evening students is known to be normally distributed with a population mean of 40 and a population standard deviation of six. A sample of six De Anza evening students reported their ages (in years) as: 28; 35; 47; 45; 30; 50. Find the probability that the mean of six ages of randomly chosen students is less than 35 years. Hint: Find the sample mean.

138. A math exam was given to all the fifth grade children attending Country School. Two random samples of scores were taken. The null hypothesis is that the mean math scores for boys and girls in fifth grade are the same. Conduct a hypothesis test.

	n	x	s^2
Boys	55	82	29
Girls	60	86	46

139. In a survey of 80 males, 55 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. Conduct a hypothesis test.

140. Which of the following is preferable when designing a hypothesis test?

a. Maximize α and minimize β
b. Minimize α and maximize β
c. Maximize α and β
d. Minimize α and β

Use the following information to answer the next three exercises: 120 people were surveyed as to their favorite beverage (non-alcoholic). The results are below.

Beverage/Age	**0–9**	**10–19**	**20–29**	**30+**	**Totals**
Milk	14	10	6	0	30
Soda	3	8	26	15	52

Beverage/Age	0–9	10–19	20–29	30+	Totals
Juice	7	12	12	7	38
Totals	24	330	44	22	120

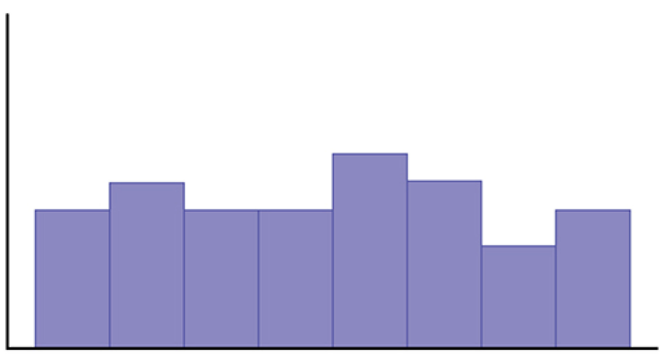
141. Are the events of milk and 30+:

a. independent events? Justify your answer.
b. mutually exclusive events? Justify your answer.

142. Suppose that one person is randomly chosen. Find the probability that person is 10–19 given that he or she prefers juice.

143. Are “Preferred Beverage” and “Age” independent events? Conduct a hypothesis test.

144. Given the following histogram, which distribution is the data most likely to come from?



a. uniform
b. exponential
c. normal
d. chi-square

Solutions

Chapter 3

1. c. parameter

2. a. population

3. b. statistic

4. d. sample

5. e. variable

6. quantitative continuous

7.

a. 2.27
b. 3.04
c. –1, 4, 4

8. Answers will vary.

Chapter 4

9. c. (0.80)(0.30)

10. b. No, and they are not mutually exclusive either.

11. a. all employed adult women

12. 0.5773

13. 0.0522

14. b. The middle fifty percent of the members lost from 2 to 8.5 lbs.

15. c. All of the data have the same value.

16. c. The lowest data value is the median.

17. 0.279

18. b. No, I expect to come out behind in money.

19. X = the number of patients calling in claiming to have the flu, who actually have the flu.
X = 0, 1, 2, …25

20. $B(25, 0.04)$

21. 0.0165

22. 1

23. c. quantitative discrete

24. all words used by Tom Clancy in his novels

Chapter 5

25.

a. 24%
b. 27%

26. qualitative

27. 0.36

28. 0.7636

29.

a. No
b. No

30. $B(10, 0.76)$

31. 0.9330

32.

a. X = the number of questions posted to the statistics listserv per day.
b. X = 0, 1, 2,…
c. $X \sim P(2)$
d. 0

33. $150

34. Matt

35.

a. false
b. true
c. false
d. false

36. 16

37. first quartile: 2
second quartile: 2
third quartile: 3

38. 0.5

39. $\frac{7}{15}$

40. $\frac{2}{15}$

Chapter 6

41.

a. true
b. true

c. False – the median and the mean are the same for this symmetric distribution.
d. true

42.

a. 8
b. 8
c. $P(x < k) = 0.65 = (k – 3)\left(\frac{1}{10}\right)$. $k = 9.5$

43.

a. False – $\frac{3}{4}$ of the data are at most five.
b. True – each quartile has 25% of the data.
c. False – that is unknown.
d. False – 50% of the data are four or less.

44. d. *G* and *H* are independent events.

45.

a. False – *J* and *K* are independent so they are not mutually exclusive which would imply dependency (meaning $P(J \text{ AND } K)$ is not 0).
b. False – see answer c.
c. True – $P(J \text{ OR } K) = P(J) + P(K) – P(J \text{ AND } K) = P(J) + P(K) – P(J)P(K) = 0.3 + 0.6 – (0.3)(0.6) = 0.72$. Note the $P(J \text{ AND } K) = P(J)P(K)$ because *J* and *K* are independent.
d. False – *J* and *K* are independent so $P(J) = P(J|K)$

46. a. $P(5)$

Chapter 7

47. a. $U(0, 4)$

48. b. 2 hour

49. a. $\frac{1}{4}$

50.

a. 0.7165
b. 4.16
c. 0

51. c. 5 years

52. c. exponential

53. 0.63

54. $B(14, 0.20)$

55. $B(14, 0.20)$

Chapter 8

56. c. the mean amount of weight lost by 15 people on the special weight loss diet.

57. 0.9951

58. 12.99

59. c. $\frac{1}{2}$

60. b. 0.60

61. c. $N(60, 5.477)$

62. 0.9990

63. a. eight days

64. c. 0.7500

65. a. 80%

66. b. 35%

67. b. no

68. b. quantitative continuous

69. c. 150

70. d. 0.06

71. c. 0.44

72. b. 0

Chapter 9

73. d. Matt is shorter than the average 14 year old boy.

74. Answers will vary.

75.

x	**Relative Frequency**	**Cumulative Relative Frequency**
1	0.3	0.3
2	0.2	0.2
4	0.4	0.4
5	0.1	0.1

76.

a. 2.8
b. 1.48
c. 90%

77. $M = 3$; $Q_1 = 1$; $Q_3 = 4$

78. 1 and 4

79. d. $\frac{8}{70}$

80. c. $\frac{40}{70}$

81. a. $\frac{9}{19}$

82. b. false

83. b. false

84. b. false

85.

a. X = the number of pies Lee bakes every day.
b. $P(20)$
c. 0.1122

86. CI: (5.25, 8.48)

87.

a. uniform
b. exponential
c. normal

Chapter 10

88. $\frac{77}{100}$

89. $\frac{12}{42}$

90.

a. false
b. false
c. true
d. false

91. $N(180, 16.43)$

92. a. The distribution for X is still uniform with the same mean and standard deviation as the distribution for X.

93. c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.

94. $N\left(2,\ \frac{0.25}{\sqrt{16}}\right)$

95. Answers will vary.

96. 0.5000

97. 7.6

98. 5

99. 0.9431

Chapter 11

100. 7.5

101. 0.0122

102. $N(7, 0.63)$

103. 0.9911

104. b. Exponential

105.

a. true
b. false
c. false

106. Answers will vary.

107. Student's t with $df = 15$

108. (560.07, 719.93)

109. quantitative continuous data

110. quantitative discrete data

111.

a. X = the number of patients with a shotgun wound the emergency room gets per 28 days
b. $P(4)$
c. 0.0183

112. greater than

113. No; $P(x = 8) = 0.0348$

114. You will lose $5.

115. Becca

116. 14

117. Sample mean = 3.2
Sample standard deviation = 1.85
Median = 3
$Q_1 = 2$
$Q_3 = 5$
$IQR = 3$

118. d. $z = –1.19$
e. 0.1171
f. Do not reject the null hypothesis.

119. We conclude that the patient does have the HIV virus when, in fact, the patient does not.

120. c. $z = 2.21$; $p = 0.0136$
d. Reject the null hypothesis.
e. We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
f. We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

121. c. dependent means

122. t_5

Chapter 12

123. (0.0424, 0.0770)

124. 2,401

125. Check student's solution.

126. 0.6321

127. $360

128. $N\left(72,\ \frac{72}{\sqrt{5}}\right)$

Chapter 13

129. 0.02

130. 0.40

131. $\frac{100}{140}$

132. $\frac{10}{60}$

133. *p*-value = 0; Reject the null hypothesis; conclude that they are dependent events

134. 8.4

135. $B(14, 0.60)$

136. d. Binomial

137. 0.3669

138. *p*-value = 0.0006; reject the null hypothesis; conclude that the averages are not equal

139. *p*-value = 0; reject the null hypothesis; conclude that the proportion of males is higher

140. Minimize α and β

141.

a. No
b. Yes, $P(M \text{ AND } 30+) = 0$

142. $\frac{12}{38}$

143. No; *p*-value = 0

144. a. uniform

References

Data from the *San Jose Mercury News*.

Baran, Daya. "20 Percent of Americans Have Never Used Email." Webguild.org, 2010. Available online at: http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed October 17, 2013).

Data from *Parade Magazine*.

Practice Tests (1-4) and Final Exams

Practice Test 1

1.1: Definitions of Statistics, Probability, and Key Terms

Use the following information to answer the next three exercises. A grocery store is interested in how much money, on average, their customers spend each visit in the produce department. Using their store records, they draw a sample of 1,000 visits and calculate each customer's average spending on produce.

1. Identify the population, sample, parameter, statistic, variable, and data for this example.

a. population
b. sample
c. parameter
d. statistic
e. variable
f. data

2. What kind of data is "amount of money spent on produce per visit"?

a. qualitative
b. quantitative-continuous
c. quantitative-discrete

3. The study finds that the mean amount spent on produce per visit by the customers in the sample is $12.84. This is an example of a:

a. population
b. sample
c. parameter
d. statistic
e. variable

1.2: Data, Sampling, and Variation in Data and Sampling

Use the following information to answer the next two exercises. A health club is interested in knowing how many times a typical member uses the club in a week. They decide to ask every tenth customer on a specified day to complete a short survey including information about how many times they have visited the club in the past week.

4. What kind of a sampling design is this?

a. cluster
b. stratified
c. simple random
d. systematic

5. “Number of visits per week” is what kind of data?

a. qualitative
b. quantitative-continuous
c. quantitative-discrete

6. Describe a situation in which you would calculate a parameter, rather than a statistic.

7. The U.S. federal government conducts a survey of high school seniors concerning their plans for future education and employment. One question asks whether they are planning to attend a four-year college or university in the following year. Fifty percent answer yes to this question; that fifty percent is a:

a. parameter
b. statistic
c. variable
d. data

8. Imagine that the U.S. federal government had the means to survey all high school seniors in the U.S. concerning their plans for future education and employment, and found that 50 percent were planning to attend a 4-year college or university in the following year. This 50 percent is an example of a:

a. parameter
b. statistic
c. variable
d. data

Use the following information to answer the next three exercises. A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table.

9. Fill in the blanks in the table and round your answers to two decimal places for the Relative Frequency and Cumulative Relative Frequency cells.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25		
5–10	30		
> 10	empty		

10. What proportion of nurses have five or more years of experience?

11. What proportion of nurses have ten or fewer years of experience?

12. Describe how you might draw a random sample of 30 students from a lecture class of 200 students.

13. Describe how you might draw a stratified sample of students from a college, where the strata are the students' class standing (freshman, sophomore, junior, or senior).

14. A manager wants to draw a sample, without replacement, of 30 employees from a workforce of 150. Describe how the chance of being selected will change over the course of drawing the sample.

15. The manager of a department store decides to measure employee satisfaction by selecting four departments at random, and conducting interviews with all the employees in those four departments. What type of survey design is this?

a. cluster
b. stratified
c. simple random
d. systematic

16. A popular American television sports program conducts a poll of viewers to see which team they believe will win the NFL (National Football League) championship this year. Viewers vote by calling a number displayed on the television screen and telling the operator which team they think will win. Do you think that those who participate in this poll are representative of all football fans in America?

17. Two researchers studying vaccination rates independently draw samples of 50 children, ages 3–18 months, from a large urban area, and determine if they are up to date on their vaccinations. One researcher finds that 84 percent of the children in her sample are up to date, and the other finds that 86 percent in his sample are up to date. Assuming both followed proper sampling procedures and did their calculations correctly, what is a likely explanation for this discrepancy?

18. A high school increased the length of the school day from 6.5 to 7.5 hours. Students who wished to attend this high school were required to sign contracts pledging to put forth their best effort on their school work and to obey the school rules; if they did not wish to do so, they could attend another high school in the district. At the end of one year, student performance on statewide tests had increased by ten percentage points over the previous year. Does this improvement prove that a longer school day improves student achievement?

19. You read a newspaper article reporting that eating almonds leads to increased life satisfaction. The study was conducted by the Almond Growers Association, and was based on a randomized survey asking people about their consumption of various foods, including almonds, and also about their satisfaction with different aspects of their life. Does anything about this poll lead you to question its conclusion?

20. Why is non-response a problem in surveys?

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. Compute the mean of the following numbers, and report your answer using one more decimal place than is present in the original data:
14, 5, 18, 23, 6

1.4: Experimental Design and Ethics

22. A psychologist is interested in whether the size of tableware (bowls, plates, etc.) influences how much college students eat. He randomly assigns 100 college students to one of two groups: the first is served a meal using normal-sized tableware, while the second is served the same meal, but using tableware that it 20 percent smaller than normal. He records how much food is consumed by each group. Identify the following components of this study.

a. population
b. sample
c. experimental units
d. explanatory variable
e. treatment
f. response variable

23. A researcher analyzes the results of the SAT (Scholastic Aptitude Test) over a five-year period and finds that male students on average score higher on the math section, and female students on average score higher on the verbal section. She concludes that these observed differences in test performance are due to genetic factors. Explain how lurking variables could offer an alternative explanation for the observed differences in test scores.

24. Explain why it would not be possible to use random assignment to study the health effects of smoking.

25. A professor conducts a telephone survey of a city's population by drawing a sample of numbers from the phone book and having her student assistants call each of the selected numbers once to administer the survey. What are some sources of bias with this survey?

26. A professor offers extra credit to students who take part in her research studies. What is an ethical problem with this method of recruiting subjects?

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Use the following information to answer the next four exercises. The midterm grades on a chemistry exam, graded on a scale of 0 to 100, were:
62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76,78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100, 740

27. Do you see any outliers in this data? If so, how would you address the situation?

28. Construct a stem plot for this data, using only the values in the range 0–100.

29. Describe the distribution of exam scores.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. In a class of 35 students, seven students received scores in the 70–79 range. What is the relative frequency of scores in this range?

Use the following information to answer the next three exercises. You conduct a poll of 30 students to see how many classes they are taking this term. Your results are:
1; 1; 1; 1
2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4; 4; 4; 4
5; 5; 5; 5

31. You decide to construct a histogram of this data. What will be the range of your first bar, and what will be the central point?

32. What will be the widths and central points of the other bars?

33. Which bar in this histogram will be the tallest, and what will be its height?

34. You get data from the U.S. Census Bureau on the median household income for your city, and decide to display it graphically. Which is the better choice for this data, a bar graph or a histogram?

35. You collect data on the color of cars driven by students in your statistics class, and want to display this information graphically. Which is the better choice for this data, a bar graph or a histogram?

2.3: Measures of the Location of the Data

36. Your daughter brings home test scores showing that she scored in the 80^{th} percentile in math and the 76^{th} percentile in reading for her grade. Interpret these scores.

37. You have to wait 90 minutes in the emergency room of a hospital before you can see a doctor. You learn that your wait time was in the 82^{nd} percentile of all wait times. Explain what this means, and whether you think it is good or bad.

2.4: Box Plots

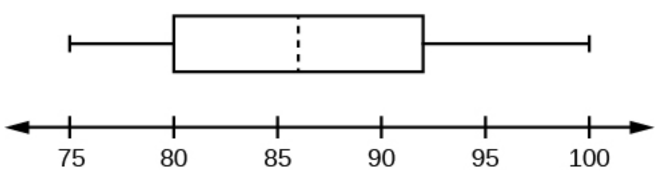
Use the following information to answer the next three exercises. 1; 1; 2; 3; 4; 4; 5; 5; 6; 7; 7; 8; 9

38. What is the median for this data?

39. What is the first quartile for this data?

40. What is the third quartile for this data?

Use the following information to answer the next four exercises. This box plot represents scores on the final exam for a physics class.



41. What is the median for this data, and how do you know?

42. What are the first and third quartiles for this data, and how do you know?

43. What is the interquartile range for this data?

44. What is the range for this data?

2.5: Measures of the Center of the Data

45. In a marathon, the median finishing time was 3:35:04 (three hours, 35 minutes, and four seconds). You finished in 3:34:10. Interpret the meaning of the median time, and discuss your time in relation to it.

Use the following information to answer the next three exercises. The value, in thousands of dollars, for houses on a block, are: 45; 47; 47.5; 51; 53.5; 125.

46. Calculate the mean for this data.

47. Calculate the median for this data.

48. Which do you think better reflects the average value of the homes on this block?

2.6: Skewness and the Mean, Median, and Mode

49. In a left-skewed distribution, which is greater?

a. the mean
b. the media
c. the mode

50. In a right-skewed distribution, which is greater?

a. the mean
b. the median
c. the mode

51. In a symmetrical distribution what will be the relationship among the mean, median, and mode?

2.7: Measures of the Spread of the Data

Use the following information to answer the next four exercises. 10; 11; 15; 15; 17; 22

52. Compute the mean and standard deviation for this data; use the sample formula for the standard deviation.

53. What number is two standard deviations above the mean of this data?

54. Express the number 13.7 in terms of the mean and standard deviation of this data.

55. In a biology class, the scores on the final exam were normally distributed, with a mean of 85, and a standard deviation of five. Susan got a final exam score of 95. Express her exam result as a z-score, and interpret its meaning.

3.1: Terminology

Use the following information to answer the next two exercises. You have a jar full of marbles: 50 are red, 25 are blue, and 15 are yellow. Assume you draw one marble at random for each trial, and replace it before the next trial.
Let $P(R)$ = the probability of drawing a red marble.
Let $P(B)$ = the probability of drawing a blue marble.
Let $P(Y)$ = the probability of drawing a yellow marble.

56. Find $P(B)$.

57. Which is more likely, drawing a red marble or a yellow marble? Justify your answer numerically.

Use the following information to answer the next two exercises. The following are probabilities describing a group of college students.
Let $P(M)$ = the probability that the student is male
Let $P(F)$ = the probability that the student is female
Let $P(E)$ = the probability the student is majoring in education
Let $P(S)$ = the probability the student is majoring in science

58. Write the symbols for the probability that a student, selected at random, is both female and a science major.

59. Write the symbols for the probability that the student is an education major, given that the student is male.

3.2: Independent and Mutually Exclusive Events

60. Events A and B are independent.
If $P(A) = 0.3$ and $P(B) = 0.5$, find $P(A \text{ AND } B)$.

61. C and D are mutually exclusive events.
If $P(C) = 0.18$ and $P(D) = 0.03$, find $P(C \text{ OR } D)$.

3.3: Two Basic Rules of Probability

62. In a high school graduating class of 300, 200 students are going to college, 40 are planning to work full-time, and 80 are taking a gap year. Are these events mutually exclusive?

Use the following information to answer the next two exercises. An archer hits the center of the target (the bullseye) 70 percent of the time. However, she is a streak shooter, and if she hits the center on one shot, her probability of hitting it on the shot immediately following is 0.85. Written in probability notation:
$P(A) = P(B) = P(\text{hitting the center on one shot}) = 0.70$
$P(B|A) = \text{P(hitting the center on a second shot, given that she hit it on the first)} = 0.85$

63. Calculate the probability that she will hit the center of the target on two consecutive shots.

64. Are $P(A)$ and $P(B)$ independent in this example?

3.4: Contingency Tables

Use the following information to answer the next three exercises. The following contingency table displays the number of students who report studying at least 15 hours per week, and how many made the honor roll in the past semester.

	Honor roll	**No honor roll**	**Total**
Study at least 15 hours/week		200	
Study less than 15 hours/week	125	193	
Total			1,000

65. Complete the table.

66. Find $P(\text{honor roll}|\text{study at least 15 hours per week})$.

67. What is the probability a student studies less than 15 hours per week?

68. Are the events “study at least 15 hours per week” and “makes the honor roll” independent? Justify your answer numerically.

3.5: Tree and Venn Diagrams

69. At a high school, some students play on the tennis team, some play on the soccer team, but neither plays both tennis and soccer. Draw a Venn diagram illustrating this.

70. At a high school, some students play tennis, some play soccer, and some play both. Draw a Venn diagram illustrating this.

Practice Test 1 Solutions

1.1: Definitions of Statistics, Probability, and Key Terms

1.

a. population: all the shopping visits by all the store’s customers
b. sample: the 1,000 visits drawn for the study
c. parameter: the average expenditure on produce per visit by all the store’s customers
d. statistic: the average expenditure on produce per visit by the sample of 1,000
e. variable: the expenditure on produce for each visit
f. data: the dollar amounts spent on produce; for instance, $15.40, $11.53, etc

2. c

3. d

1.2: Data, Sampling, and Variation in Data and Sampling

4. d

5. c

6. Answers will vary.
Sample Answer: Any solution in which you use data from the entire population is acceptable. For instance, a professor might calculate the average exam score for her class: because the scores of all members of the class were used in the calculation, the average is a parameter.

7. b

8. a

9.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25	0.25	0.25

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
5–10	30	0.30	0.55
> 10	45	0.45	1.00

10. 0.75

11. 0.55

12. Answers will vary.
Sample Answer: One possibility is to obtain the class roster and assign each student a number from 1 to 200. Then use a random number generator or table of random number to generate 30 numbers between 1 and 200, and select the students matching the random numbers. It would also be acceptable to write each student's name on a card, shuffle them in a box, and draw 30 names at random.

13. One possibility would be to obtain a roster of students enrolled in the college, including the class standing for each student. Then you would draw a proportionate random sample from within each class (for instance, if 30 percent of the students in the college are freshman, then 30 percent of your sample would be drawn from the freshman class).

14. For the first person picked, the chance of any individual being selected is one in 150. For the second person, it is one in 149, for the third it is one in 148, and so on. For the 30th person selected, the chance of selection is one in 121.

15. a

16. No. There are at least two chances for bias. First, the viewers of this particular program may not be representative of American football fans as a whole. Second, the sample will be self-selected, because people have to make a phone call in order to take part, and those people are probably not representative of the American football fan population as a whole.

17. These results (84 percent in one sample, 86 percent in the other) are probably due to sampling variability. Each researcher drew a different sample of children, and you would not expect them to get exactly the same result, although you would expect the results to be similar, as they are in this case.

18. No. The improvement could also be due to self-selection: only motivated students were willing to sign the contract, and they would have done well even in a school with 6.5 hour days. Because both changes were implemented at the same time, it is not possible to separate out their influence.

19. At least two aspects of this poll are troublesome. The first is that it was conducted by a group who would benefit by the result—almond sales are likely to increase if people believe that eating almonds will make them happier. The second is that this poll found that almond consumption and life satisfaction are correlated, but does not establish that eating almonds causes satisfaction. It is equally possible, for instance, that people with higher incomes are more likely to eat almonds, and are also more satisfied with their lives.

20. You want the sample of people who take part in a survey to be representative of the population from which they are drawn. People who refuse to take part in a survey often have different views than those who do participate, and so even a random sample may produce biased results if a large percentage of those selected refuse to participate in a survey.

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. 13.2

1.4: Experimental Design and Ethics

22.

a. population: all college students
b. sample: the 100 college students in the study
c. experimental units: each individual college student who participated
d. explanatory variable: the size of the tableware
e. treatment: tableware that is 20 percent smaller than normal
f. response variable: the amount of food eaten

23. There are many lurking variables that could influence the observed differences in test scores. Perhaps the boys, on average, have taken more math courses than the girls, and the girls have taken more English classes than the boys. Perhaps the boys have been encouraged by their families and teachers to prepare for a career in math and science, and thus have put more effort into studying math, while the girls have been encouraged to prepare for fields like communication and psychology that are more focused on language use. A study design would have to control for these and other potential lurking variables (anything that could explain the observed difference in test scores, other than the genetic explanation) in order to draw a scientifically sound conclusion about genetic differences.

24. To use random assignment, you would have to be able to assign people to either smoke or not smoke. Because smoking has many harmful effects, this would not be an ethical experiment. Instead, we study people who have chosen to smoke, and compare them to others who have chosen not to smoke, and try to control for the other ways those two groups may differ (lurking variables).

25. Sources of bias include the fact that not everyone has a telephone, that cell phone numbers are often not listed in published directories, and that an individual might not be at home at the time of the phone call; all these factors make it likely that the respondents to the survey will not be representative of the population as a whole.

26. Research subjects should not be coerced into participation, and offering extra credit in exchange for participation could be construed as coercion. In addition, this method will result in a volunteer sample, which cannot be assumed to be representative of the population as a whole.

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

27. The value 740 is an outlier, because the exams were graded on a scale of 0 to 100, and 740 is far outside that range. It may be a data entry error, with the actual score being 74, so the professor should check that exam again to see what the actual score was.

28.

Stem	Leaf
6	2 4 5 5 8
7	0 2 2 4 5 5 5 6 8 8
8	1 3 3 4 5 7 8
9	2 5 8 8

Stem	Leaf
10	0 0

29. Most scores on this exam were in the range of 70–89, with a few scoring in the 60–69 range, and a few in the 90–100 range.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. $RF = \frac{7}{35} = 0.2$

31. The range will be 0.5–1.5, and the central point will be 1.

32. Range 1.5–2.5, central point 2; range 2.5–3.5, central point 3; range 3.5–4.5, central point 4; range 4.5–5.5., central point 5.

33. The bar from 3.5 to 4.5, with a central point of 4, will be tallest; its height will be nine, because there are nine students taking four courses.

34. The histogram is a better choice, because income is a continuous variable.

35. A bar graph is the better choice, because this data is categorical rather than continuous.

2.3: Measures of the Location of the Data

36. Your daughter scored better than 80 percent of the students in her grade on math and better than 76 percent of the students in reading. Both scores are very good, and place her in the upper quartile, but her math score is slightly better in relation to her peers than her reading score.

37. You had an unusually long wait time, which is bad: 82 percent of patients had a shorter wait time than you, and only 18 percent had a longer wait time.

2.4: Box Plots

38. 5

39. 3

40. 7

41. The median is 86, as represented by the vertical line in the box.

42. The first quartile is 80, and the third quartile is 92, as represented by the left and right boundaries of the box.

43. IQR = 92 – 80 = 12

44. Range = 100 – 75 = 25

2.5: Measures of the Center of the Data

45. Half the runners who finished the marathon ran a time faster than 3:35:04, and half ran a time slower than 3:35:04. Your time is faster than the median time, so you did better than more than half of the runners in this race.

46. 61.5, or $61,500

47. 49.25 or $49,250

48. The median, because the mean is distorted by the high value of one house.

2.6: Skewness and the Mean, Median, and Mode

49. c

50. a

51. They will all be fairly close to each other.

2.7: Measures of the Spread of the Data

52. Mean: 15
Standard deviation: 4.3

$\mu = \frac{10+11+15+15+17+22}{6} = 15$

$s = \sqrt{\frac{\sum (x-x)^2}{n-1}} = \sqrt{\frac{94}{5}} = 4.3$

53. 15 + (2)(4.3) = 23.6

54. 13.7 is one standard deviation below the mean of this data, because 15 – 4.3 = 10.7

55. $z = \frac{95-85}{5} = 2.0$
Susan's *z*-score was 2.0, meaning she scored two standard deviations above the class mean for the final exam.

3.1: Terminology

56. $P(B) = \frac{25}{90} = 0.28$

57. Drawing a red marble is more likely.
$P(R) = \frac{50}{80} = 0.62$
$P(Y) = \frac{15}{80} = 0.19$

58. $P(F \text{ AND } S)$

59. $P(E|M)$

3.2: Independent and Mutually Exclusive Events

60. $P(A \text{ AND } B)$ = (0.3)(0.5) = 0.15

61. $P(C \text{ OR } D) = 0.18 + 0.03 = 0.21$

3.3: Two Basic Rules of Probability

62. No, they cannot be mutually exclusive, because they add up to more than 300. Therefore, some students must fit into two or more categories (e.g., both going to college and working full time).

63. $P(A \text{ and } B) = (P(B|A))(P(A)) = (0.85)(0.70) = 0.595$

64. No. If they were independent, $P(B)$ would be the same as $P(B|A)$. We know this is not the case, because $P(B) = 0.70$ and $P(B|A) = 0.85$.

3.4: Contingency Tables

65.

	Honor roll	No honor roll	Total
Study at least 15 hours/week	482	200	682
Study less than 15 hours/week	125	193	318
Total	607	393	1,000

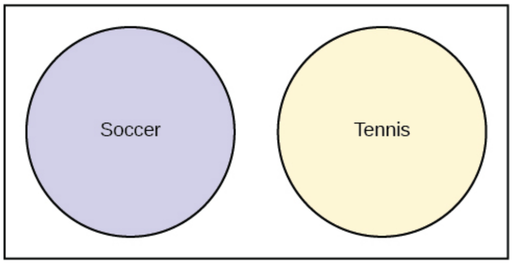
66. $P(\text{honor roll}|\text{study at least 15 hours word per week}) = \frac{482}{1000} = 0.482$

67. $P(\text{studies less than 15 hours word per week}) = \frac{125+193}{1000} = 0.318$

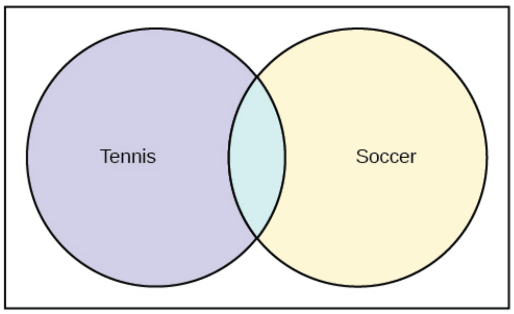
68. Let $P(S)$ = study at least 15 hours per week
Let $P(H)$ = makes the honor roll
From the table, $P(S) = 0.682$, $P(H) = 0.607$, and $P(S \text{ AND } H) = 0.482$.
If $P(S)$ and $P(H)$ were independent, then $P(S \text{ AND } H)$ would equal $(P(S))(P(H))$.
However, $(P(S))(P(H)) = (0.682)(0.607) = 0.414$, while $P(S \text{ AND } H) = 0.482$.
Therefore, $P(S)$ and $P(H)$ are not independent.

3.5: Tree and Venn Diagrams

69.



70.



Practice Test 2

4.1: Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises. You conduct a survey among a random sample of students at a particular university. The data collected includes their major, the number of classes they took the previous semester, and amount of money they spent on books purchased for classes in the previous semester.

1. If X = student's major, then what is the domain of X?

2. If Y = the number of classes taken in the previous semester, what is the domain of Y?

3. If Z = the amount of money spent on books in the previous semester, what is the domain of Z?

4. Why are X, Y, and Z in the previous example random variables?

5. After collecting data, you find that for one case, z = –7. Is this a possible value for Z?

6. What are the two essential characteristics of a discrete probability distribution?

Use this discrete probability distribution represented in this table to answer the following six questions. The university library records the number of books checked out by each patron over the course of one day, with the following result:

x	***P*(*x*)**

x	$P(x)$
0	0.20
1	0.45
2	0.20
3	0.10
4	0.05

7. Define the random variable X for this example.

8. What is $P(x > 2)$?

9. What is the probability that a patron will check out at least one book?

10. What is the probability a patron will take out no more than three books?

11. If the table listed $P(x)$ as 0.15, how would you know that there was a mistake?

12. What is the average number of books taken out by a patron?

4.2: Mean or Expected Value and Standard Deviation

Use the following information to answer the next four exercises. Three jobs are open in a company: one in the accounting department, one in the human resources department, and one in the sales department. The accounting job receives 30 applicants, and the human resources and sales department 60 applicants.

13. If X = the number of applications for a job, use this information to fill in [link].

x	$P(x)$	$xP(x)$

14. What is the mean number of applicants?

15. What is the PDF for X?

16. Add a fourth column to the table, for $(x – \mu)^2P(x)$.

17. What is the standard deviation of X?

4.3: Binomial Distribution

18. In a binomial experiment, if $p = 0.65$, what does q equal?

19. What are the required characteristics of a binomial experiment?

20. Joe conducts an experiment to see how many times he has to flip a coin before he gets four heads in a row. Does this qualify as a binomial experiment?

Use the following information to answer the next three exercises. In a particularly community, 65 percent of households include at least one person who has graduated from college. You randomly sample 100 households in this community. Let X = the number of households including at least one college graduate.

21. Describe the probability distribution of X.

22. What is the mean of X?

23. What is the standard deviation of X?

Use the following information to answer the next four exercises. Joe is the star of his school's baseball team. His batting average is 0.400, meaning that for every ten times he comes to bat (an at-bat), four of those times he gets a hit. You decide to track his batting performance his next 20 at-bats.

24. Define the random variable X in this experiment.

25. Assuming Joe's probability of getting a hit is independent and identical across all 20 at-bats, describe the distribution of X.

26. Given this information, what number of hits do you predict Joe will get?

27. What is the standard deviation of X?

4.4: Geometric Distribution

28. What are the three major characteristics of a geometric experiment?

29. You decide to conduct a geometric experiment by flipping a coin until it comes up heads. This takes five trials. Represent the outcomes of this trial, using H for heads and T for tails.

30. You are conducting a geometric experiment by drawing cards from a normal 52-card pack, with replacement, until you draw the Queen of Hearts. What is the domain of X for this experiment?

31. You are conducting a geometric experiment by drawing cards from a normal 52-card deck, without replacement, until you draw a red card. What is the domain of X for this experiment?

Use the following information to answer the next three exercises. In a particular university, 27 percent of students are engineering majors. You decide to select students at random until you choose one that is an engineering major. Let X = the number of students you select until you find one that is an engineering major.

32. What is the probability distribution of X?

33. What is the mean of X?

34. What is the standard deviation of X?

4.5: Hypergeometric Distribution

35. You draw a random sample of ten students to participate in a survey, from a group of 30, consisting of 16 boys and 14 girls. You are interested in the probability that seven of the students chosen will be boys. Does this qualify as a hypergeometric experiment? List the conditions and whether or not they are met.

36. You draw five cards, without replacement, from a normal 52-card deck of playing cards, and are interested in the probability that two of the cards are spades. What are the group of interest, size of the group of interest, and sample size for this example?

4.6: Poisson Distribution

37. What are the key characteristics of the Poisson distribution?

Use the following information to answer the next three exercises. The number of drivers to arrive at a toll booth in an hour can be modeled by the Poisson distribution.

38. If X = the number of drivers, and the average numbers of drivers per hour is four, how would you express this distribution?

39. What is the domain of X?

40. What are the mean and standard deviation of X?

5.1: Continuous Probability Functions

41. You conduct a survey of students to see how many books they purchased the previous semester, the total amount they paid for those books, the number they sold after the semester was over, and the amount of money they received for the books they sold. Which variables in this survey are discrete, and which are continuous?

42. With continuous random variables, we never calculate the probability that X has a particular value, but always speak in terms of the probability that X has a value within a particular range. Why is this?

43. For a continuous random variable, why are $P(x < c)$ and $P(x \leq c)$ equivalent statements?

44. For a continuous probability function, $P(x < 5) = 0.35$. What is $P(x > 5)$, and how do you know?

45. Describe how you would draw the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. What type of a distribution is this?

46. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(0 < x < 4)$?

5.2: The Uniform Distribution

47. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(2 < x < 5)$?

Use the following information to answer the next four exercises. The number of minutes that a patient waits at a medical clinic to see a doctor is represented by a uniform distribution between zero and 30 minutes, inclusive.

48. If X equals the number of minutes a person waits, what is the distribution of X?

49. Write the probability density function for this distribution.

50. What is the mean and standard deviation for waiting time?

51. What is the probability that a patient waits less than ten minutes?

5.3: The Exponential Distribution

52. The distribution of the variable *X*, representing the average time to failure for an automobile battery, can be written as: *X* ~ *Exp*(*m*). Describe this distribution in words.

53. If the value of *m* for an exponential distribution is ten, what are the mean and standard deviation for the distribution?

54. Write the probability density function for a variable distributed as: *X* ~ *Exp*(0.2).

6.1: The Standard Normal Distribution

55. Translate this statement about the distribution of a random variable *X* into words: *X* ~ (100, 15).

56. If the variable *X* has the standard normal distribution, express this symbolically.

Use the following information for the next six exercises. According to the World Health Organization, distribution of height in centimeters for girls aged five years and no months has the distribution: *X* ~ *N*(109, 4.5).

57. What is the *z*-score for a height of 112 inches?

58. What is the *z*-score for a height of 100 centimeters?

59. Find the *z*-score for a height of 105 centimeters and explain what that means In the context of the population.

60. What height corresponds to a *z*-score of 1.5 in this population?

61. Using the empirical rule, we expect about 68 percent of the values in a normal distribution to lie within one standard deviation above or below the mean. What does this mean, in terms of a specific range of values, for this distribution?

62. Using the empirical rule, about what percent of heights in this distribution do you expect to be between 95.5 cm and 122.5 cm?

6.2: Using the Normal Distribution

Use the following information to answer the next four exercises. The distributor of lotto tickets claims that 20 percent of the tickets are winners. You draw a sample of 500 tickets to test this proposition.

63. Can you use the normal approximation to the binomial for your calculations? Why or why not.

64. What are the expected mean and standard deviation for your sample, assuming the distributor's claim is true?

65. What is the probability that your sample will have a mean greater than 100?

66. If the *z*-score for your sample result is –2.00, explain what this means, using the empirical rule.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. What does the central limit theorem state with regard to the distribution of sample means?

68. The distribution of results from flipping a fair coin is uniform: heads and tails are equally likely on any flip, and over a large number of trials, you expect about the same number of heads and tails. Yet if you conduct a study by flipping 30 coins and recording the number of heads, and repeat this 100 times, the distribution of the mean number of heads will be approximately normal. How is this possible?

69. The mean of a normally-distributed population is 50, and the standard deviation is four. If you draw 100 samples of size 40 from this population, describe what you would expect to see in terms of the sampling distribution of the sample mean.

70. X is a random variable with a mean of 25 and a standard deviation of two. Write the distribution for the sample mean of samples of size 100 drawn from this population.

71. Your friend is doing an experiment drawing samples of size 50 from a population with a mean of 117 and a standard deviation of 16. This sample size is large enough to allow use of the central limit theorem, so he says the standard deviation of the sampling distribution of sample means will also be 16. Explain why this is wrong, and calculate the correct value.

72. You are reading a research article that refers to "the standard error of the mean." What does this mean, and how is it calculated?

Use the following information to answer the next six exercises. You repeatedly draw samples of $n = 100$ from a population with a mean of 75 and a standard deviation of 4.5.

73. What is the expected distribution of the sample means?

74. One of your friends tries to convince you that the standard error of the mean should be 4.5. Explain what error your friend made.

75. What is the z-score for a sample mean of 76?

76. What is the z-score for a sample mean of 74.7?

77. What sample mean corresponds to a z-score of 1.5?

78. If you decrease the sample size to 50, will the standard error of the mean be smaller or larger? What would be its value?

Use the following information to answer the next two questions. We use the empirical rule to analyze data for samples of size 60 drawn from a population with a mean of 70 and a standard deviation of 9.

79. What range of values would you expect to include 68 percent of the sample means?

80. If you increased the sample size to 100, what range would you expect to contain 68 percent of the sample means, applying the empirical rule?

7.2: The Central Limit Theorem for Sums

81. How does the central limit theorem apply to sums of random variables?

82. Explain how the rules applying the central limit theorem to sample means, and to sums of a random variable, are similar.

83. If you repeatedly draw samples of size 50 from a population with a mean of 80 and a standard deviation of four, and calculate the sum of each sample, what is the expected distribution of these sums?

Use the following information to answer the next four exercises. You draw one sample of size 40 from a population with a mean of 125 and a standard deviation of seven.

84. Compute the sum. What is the probability that the sum for your sample will be less than 5,000?

85. If you drew samples of this size repeatedly, computing the sum each time, what range of values would you expect to contain 95 percent of the sample sums?

86. What value is one standard deviation below the mean?

87. What value corresponds to a z-score of 2.2?

7.3: Using the Central Limit Theorem

88. What does the law of large numbers say about the relationship between the sample mean and the population mean?

89. Applying the law of large numbers, which sample mean would expect to be closer to the population mean, a sample of size ten or a sample of size 100?

Use this information for the next three questions. A manufacturer makes screws with a mean diameter of 0.15 cm (centimeters) and a range of 0.10 cm to 0.20 cm; within that range, the distribution is uniform.

90. If X = the diameter of one screw, what is the distribution of X?

91. Suppose you repeatedly draw samples of size 100 and calculate their mean. Applying the central limit theorem, what is the distribution of these sample means?

92. Suppose you repeatedly draw samples of 60 and calculate their sum. Applying the central limit theorem, what is the distribution of these sample sums?

Practice Test 2 Solutions

Probability Distribution Function (PDF) for a Discrete Random Variable

1. The domain of X = {English, Mathematics,….], i.e., a list of all the majors offered at the university, plus "undeclared."

2. The domain of Y = {0, 1, 2, …}, i.e., the integers from 0 to the upper limit of classes allowed by the university.

3. The domain of Z = any amount of money from 0 upwards.

4. Because they can take any value within their domain, and their value for any particular case is not known until the survey is completed.

5. No, because the domain of Z includes only positive numbers (you can't spend a negative amount of money). Possibly the value –7 is a data entry error, or a special code to indicated that the student did not answer the question.

6. The probabilities must sum to 1.0, and the probabilities of each event must be between 0 and 1, inclusive.

7. Let X = the number of books checked out by a patron.

8. $P(x > 2) = 0.10 + 0.05 = 0.15$

9. $P(x \geq 0) = 1 – 0.20 = 0.80$

10. $P(x \le 3) = 1 - 0.05 = 0.95$

11. The probabilities would sum to 1.10, and the total probability in a distribution must always equal 1.0.

12. x = 0(0.20) + 1(0.45) + 2(0.20) + 3(0.10) + 4(0.05) = 1.35

Mean or Expected Value and Standard Deviation

13.

x	***P*(*x*)**	***xP*(*x*)**
30	0.33	9.90
40	0.33	13.20
60	0.33	19.80

14. x = 9.90 + 13.20 + 19.80 = 42.90

15. $P(x = 30) = 0.33$
$P(x = 40) = 0.33$
$P(x = 60) = 0.33$

16.

x	***P*(*x*)**	***xP*(*x*)**	**$(x - \mu)^2 P(x)$**
30	0.33	9.90	$(30 - 42.90)^2(0.33) = 54.91$
40	0.33	13.20	$(40 - 42.90)^2(0.33) = 2.78$
60	0.33	19.90	$(60 - 42.90)^2(0.33) = 96.49$

17. $\sigma_x = \sqrt{54.91 + 2.78 + 96.49} = 12.42$

Binomial Distribution

18. $q = 1 - 0.65 = 0.35$

19.

1. There are a fixed number of trials.
2. There are only two possible outcomes, and they add up to 1.
3. The trials are independent and conducted under identical conditions.

20. No, because there are not a fixed number of trials

21. *X* ~ *B*(100, 0.65)

22. *μ* = *np* = 100(0.65) = 65

23. $\sigma_x = \sqrt{npq} = \sqrt{100(0.65)(0.35)} = 4.77$

24. *X* = Joe gets a hit in one at-bat (in one occasion of his coming to bat)

25. *X* ~ *B*(20, 0.4)

26. *μ* = *np* = 20(0.4) = 8

27.$\sigma_x = \sqrt{npq} = \sqrt{20(0.40)(0.60)} = 2.19$

4.4: Geometric Distribution

28.

1. A series of Bernoulli trials are conducted until one is a success, and then the experiment stops.
2. At least one trial is conducted, but there is no upper limit to the number of trials.
3. The probability of success or failure is the same for each trial.

29. *T T T T H*

30. The domain of *X* = {1, 2, 3, 4, 5, ….n}. Because you are drawing with replacement, there is no upper bound to the number of draws that may be necessary.

31. The domain of *X* = {1, 2, 3, 4, 5, 6, 7, 8., 9, 10, 11, 12…27}. Because you are drawing without replacement, and 26 of the 52 cards are red, you have to draw a red card within the first 17 draws.

32. *X* ~ *G*(0.24)

33. $\mu = \frac{1}{p} = \frac{1}{0.27} = 3.70$

34. $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.27}{0.27^2}} = 3.16$

4.5: Hypergeometric Distribution

35. Yes, because you are sampling from a population composed of two groups (boys and girls), have a group of interest (boys), and are sampling without replacement (hence, the probabilities change with each pick, and you are not performing Bernoulli trials).

36. The group of interest is the cards that are spades, the size of the group of interest is 13, and the sample size is five.

4.6: Poisson Distribution

37. A Poisson distribution models the number of events occurring in a fixed interval of time or space, when the events are independent and the average rate of the events is known.

38. $X \sim P(4)$

39. The domain of X = {0, 1, 2, 3, …..) i.e., any integer from 0 upwards.

40. $\mu = 4$
$\sigma = \sqrt{4} = 2$

5.1: Continuous Probability Functions

41. The discrete variables are the number of books purchased, and the number of books sold after the end of the semester. The continuous variables are the amount of money spent for the books, and the amount of money received when they were sold.

42. Because for a continuous random variable, $P(x = c) = 0$, where c is any single value. Instead, we calculate $P(c < x < d)$, i.e., the probability that the value of x is between the values c and d.

43. Because $P(x = c) = 0$ for any continuous random variable.

44. $P(x > 5) = 1 – 0.35 = 0.65$, because the total probability of a continuous probability function is always 1.

45. This is a uniform probability distribution. You would draw it as a rectangle with the vertical sides at 0 and 20, and the horizontal sides at $\frac{1}{10}$ and 0.

46. $P\left(0\ < x < 4\right) = \left(4-0\right)\left(\frac{1}{10}\right) =\ 0.4$

5.2: The Uniform Distribution

47. $P\left(2\ < x < 5\right) = \left(5-2\right)\left(\frac{1}{10}\right) =\ 0.3$

48. $X \sim U(0, 15)$

49. $f(x) = \frac{1}{b-a}$ for $(a \leq x \leq b)$ so $f(x) = \frac{1}{30}$ for $(0 \leq x \leq 30)$

50. $\mu = \ \frac{a+b}{2}\ =\ \frac{0+30}{5}\ = 15.0$

$\sigma =\ \sqrt{\frac{(b-a)^2}{12}}\ =\ \sqrt{\frac{(30-0)^2}{12}} = 8.66$

51. $P\left(x < 10\right) = \left(10\right)\left(\frac{1}{30}\right) =\ 0.33$

5.3: The Exponential Distribution

52. X has an exponential distribution with decay parameter m and mean and standard deviation $\frac{1}{m}$. In this distribution, there will be a relatively large numbers of small values, with values becoming less common as they become larger.

53. $\mu = \sigma = \frac{1}{m} = \frac{1}{10} = 0.1$

54. $f(x) = 0.2e^{-0.2x}$ where $x \geq 0$.

6.1: The Standard Normal Distribution

55. The random variable *X* has a normal distribution with a mean of 100 and a standard deviation of 15.

56. *X* ~ *N*(0,1)

57. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{112-109}{4.5} = 0.67$

58. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{100-109}{4.5} = -2.00$

59. $z = \frac{105-109}{4.5} = -0.89$
This girl is shorter than average for her age, by 0.89 standard deviations.

60. 109 + (1.5)(4.5) = 115.75 cm

61. We expect about 68 percent of the heights of girls of age five years and zero months to be between 104.5 cm and 113.5 cm.

62. We expect 99.7 percent of the heights in this distribution to be between 95.5 cm and 122.5 cm, because that range represents the values three standard deviations above and below the mean.

6.2: Using the Normal Distribution

63. Yes, because both np and nq are greater than five.
np = (500)(0.20) = 100 and *nq* = 500(0.80) = 400

64. $\mu = np = (500)(0.20) = 100$

$\sigma = \sqrt{npq} = \sqrt{500(0.20)(0.80)} = 8.94$

65. Fifty percent, because in a normal distribution, half the values lie above the mean.

66. The results of our sample were two standard deviations below the mean, suggesting it is unlikely that 20 percent of the lotto tickets are winners, as claimed by the distributor, and that the true percent of winners is lower. Applying the Empirical Rule, If that claim were true, we would expect to see a result this far below the mean only about 2.5 percent of the time.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. The central limit theorem states that if samples of sufficient size drawn from a population, the distribution of sample means will be normal, even if the distribution of the population is not normal.

68. The sample size of 30 is sufficiently large in this example to apply the central limit theorem. This theorem] states that for samples of sufficient size drawn from a population, the sampling distribution of the sample mean will approach normality, regardless of the distribution of the population from which the samples were drawn.

69. You would not expect each sample to have a mean of 50, because of sampling variability. However, you would expect the sampling distribution of the sample means to cluster around 50, with an approximately normal distribution, so that values close to 50 are more common than values further removed from 50.

70. $X \sim N(25, 0.2)$ because $X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

71. The standard deviation of the sampling distribution of the sample means can be calculated using the formula $\left(\frac{\sigma_x}{\sqrt{n}}\right)$, which in this case is $\left(\frac{16}{\sqrt{50}}\right)$. The correct value for the standard deviation of the sampling distribution of the sample means is therefore 2.26.

72. The standard error of the mean is another name for the standard deviation of the sampling distribution of the sample mean. Given samples of size n drawn from a population with standard deviation σ_x, the standard error of the mean is $\left(\frac{\sigma_x}{\sqrt{n}}\right)$.

73. $X \sim N(75, 0.45)$

74. Your friend forgot to divide the standard deviation by the square root of *n*.

75. $z = \frac{x - \mu_x}{\sigma_x} = \frac{76-75}{4.5} = 2.2$

76. $z = \frac{x - \mu_x}{\sigma_x} = \frac{74.7-75}{4.5} = -0.67$

77. 75 + (1.5)(0.45) = 75.675

78. The standard error of the mean will be larger, because you will be dividing by a smaller number. The standard error of the mean for samples of size *n* = 50 is:
$\left(\frac{\sigma_x}{\sqrt{n}}\right) = \frac{4.5}{\sqrt{50}} = 0.64$

79. You would expect this range to include values up to one standard deviation above or below the mean of the sample means. In this case:
$70 + \frac{9}{\sqrt{60}} = 71.16$ and $70 - \frac{9}{\sqrt{60}} = 68.84$ so you would expect 68 percent of the sample means to be between 68.84 and 71.16.

80. $70 + \frac{9}{\sqrt{100}} = 70.9$ and $70 - \frac{9}{\sqrt{100}} = 69.1$ so you would expect 68 percent of the sample means to be between 69.1 and 70.9. Note that this is a narrower interval due to the increased sample size.

7.2: The Central Limit Theorem for Sums

81. For a random variable *X*, the random variable *ΣX* will tend to become normally distributed as the size n of the samples used to compute the sum increases.

82. Both rules state that the distribution of a quantity (the mean or the sum) calculated on samples drawn from a population will tend to have a normal distribution, as the sample size increases, regardless of the distribution of population from which the samples are drawn.

83. $\Sigma X \sim N\left(n\mu_x, \left(\sqrt{n}\right)(\sigma_x)\right)$ so $\Sigma X \sim N(4000, 28.3)$

84.The probability is 0.50, because 5,000 is the mean of the sampling distribution of sums of size 40 from this population. Sums of random variables computed from a sample of sufficient size are normally distributed, and in a normal distribution, half the values lie below the mean.

85. Using the empirical rule, you would expect 95 percent of the values to be within two standard deviations of the mean. Using the formula for the standard deviation is for a sample sum: $\left(\sqrt{n}\right)(\sigma_x) = \left(\sqrt{40}\right)(7) = 44.3$ so you would expect 95 percent of the values to be between 5,000 + (2)(44.3) and 5,000 – (2)(44.3), or between 4,911.4 and 588.6.

86. $\mu - \left(\sqrt{n}\right)(\sigma_x) = 5000 - \left(\sqrt{40}\right)(7) = 4955.7$

87. $5000+(2.2)\left(\sqrt{40}\right)(7)=5097.4$

7.3: Using the Central Limit Theorem

88. The law of large numbers says that as sample size increases, the sample mean tends to get nearer and nearer to the population mean.

89. You would expect the mean from a sample of size 100 to be nearer to the population mean, because the law of large numbers says that as sample size increases, the sample mean tends to approach the population mea.

90. $X \sim N(0.10, 0.20)$

91. $X\sim N\left(\mu_x,\frac{\sigma_x}{\sqrt{n}}\right)$ and the standard deviation of a uniform distribution is $\frac{b-a}{\sqrt{12}}$. In this example, the standard deviation of the distribution is $\frac{b-a}{\sqrt{12}}=\frac{0.10}{\sqrt{12}}=0.03$
so $X\sim N\left(0.15,0.003\right)$

92. $\Sigma X\sim N((n)(\mu_x),(\sqrt{n})(\sigma_x))$ so $\Sigma X\sim N(9.0,0.23)$

Practice Test 3

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Use the following information to answer the next seven exercises. You draw a sample of size 30 from a normally distributed population with a standard deviation of four.

1. What is the standard error of the sample mean in this scenario, rounded to two decimal places?

2. What is the distribution of the sample mean?

3. If you want to construct a two-sided 95% confidence interval, how much probability will be in each tail of the distribution?

4. What is the appropriate *z*-score and error bound or margin of error (*EBM*) for a 95% confidence interval for this data?

5. Rounding to two decimal places, what is the 95% confidence interval if the sample mean is 41?

6. What is the 90% confidence interval if the sample mean is 41? Round to two decimal places

7. Suppose the sample size in this study had been 50, rather than 30. What would the 95% confidence interval be if the sample mean is 41? Round your answer to two decimal places.

8. For any given data set and sampling situation, which would you expect to be wider: a 95% confidence interval or a 99% confidence interval?

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. Comparing graphs of the standard normal distribution (*z*-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how do they differ?

10. Comparing graphs of the standard normal distribution (*z*-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how are they similar?

Use the following information to answer the next five exercises. Body temperature is known to be distributed normally among healthy adults. Because you do not know the population standard deviation, you use the t-distribution to study body temperature. You collect data from a random sample of 20 healthy adults and find that your sample temperatures have a mean of 98.4 and a sample standard deviation of 0.3 (both in degrees Fahrenheit).

11. What is the degrees of freedom (*df*) for this study?

12. For a two-tailed 95% confidence interval, what is the appropriate *t*-value to use in the formula?

13. What is the 95% confidence interval?

14. What is the 99% confidence interval? Round to two decimal places.

15. Suppose your sample size had been 30 rather than 20. What would the 95% confidence interval be then? Round to two decimal places

8.3: Confidence Interval for a Population Proportion

Use this information to answer the next four exercises. You conduct a poll of 500 randomly selected city residents, asking them if they own an automobile. 280 say they do own an automobile, and 220 say they do not.

16. Find the sample proportion and sample standard deviation for this data.

17. What is the 95% two-sided confidence interval? Round to four decimal places.

18. Calculate the 90% confidence interval. Round to four decimal places.

19. Calculate the 99% confidence interval. Round to four decimal places.

Use the following information to answer the next three exercises. You are planning to conduct a poll of community members age 65 and older, to determine how many own mobile phones. You want to produce an estimate whose 95% confidence interval will be within four percentage points (plus or minus) the true population proportion. Use an estimated population proportion of 0.5.

20. What sample size do you need?

21. Suppose you knew from prior research that the population proportion was 0.6. What sample size would you need?

22. Suppose you wanted a 95% confidence interval within three percentage points of the population. Assume the population proportion is 0.5. What sample size do you need?

9.1: Null and Alternate Hypotheses

23. In your state, 58 percent of registered voters in a community are registered as Republicans. You want to conduct a study to see if this also holds up in your community. State the null and alternative hypotheses to test this.

24. You believe that at least 58 percent of registered voters in a community are registered as Republicans. State the null and alternative hypotheses to test this.

25. The mean household value in a city is $268,000. You believe that the mean household value in a particular neighborhood is lower than the city average. Write the null and alternative hypotheses to test this.

26. State the appropriate alternative hypothesis to this null hypothesis: H_0: $\mu = 107$

27. State the appropriate alternative hypothesis to this null hypothesis: H_0: $p < 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. If you reject H_0 when H_0 is correct, what type of error is this?

29. If you fail to reject H_0 when H_0 is false, what type of error is this?

30. What is the relationship between the Type II error and the power of a test?

31. A new blood test is being developed to screen patients for cancer. Positive results are followed up by a more accurate (and expensive) test. It is assumed that the patient does not have cancer. Describe the null hypothesis, the Type I and Type II errors for this situation, and explain which type of error is more serious.

32. Explain in words what it means that a screening test for TB has an α level of 0.10. The null hypothesis is that the patient does not have TB.

33. Explain in words what it means that a screening test for TB has a β level of 0.20. The null hypothesis is that the patient does not have TB.

34. Explain in words what it means that a screening test for TB has a power of 0.80.

9.3: Distribution Needed for Hypothesis Testing

35. If you are conducting a hypothesis test of a single population mean, and you do not know the population variance, what test will you use if the sample size is 10 and the population is normal?

36. If you are conducting a hypothesis test of a single population mean, and you know the population variance, what test will you use?

37. If you are conducting a hypothesis test of a single population proportion, with *np* and *nq* greater than or equal to five, what test will you use, and with what parameters?

38. Published information indicates that, on average, college students spend less than 20 hours studying per week. You draw a sample of 25 students from your college, and find the sample mean to be 18.5 hours, with a standard deviation of 1.5 hours. What distribution will you use to test whether study habits at your college are the same as the national average, and why?

39. A published study says that 95 percent of American children are vaccinated against measles, with a standard deviation of 1.5 percent. You draw a sample of 100 children from your community and check their vaccination records, to see if the vaccination rate in your community is the same as the national average. What distribution will you use for this test, and why?

9.4: Rare Events, the Sample, Decision, and Conclusion

40. You are conducting a study with an α level of 0.05. If you get a result with a p-value of 0.07, what will be your decision?

41. You are conducting a study with $\alpha = 0.01$. If you get a result with a *p*-value of 0.006, what will be your decision?

Use the following information to answer the next five exercises. According to the World Health Organization, the average height of a one-year-old child is 29”. You believe children with a particular disease are smaller than

average, so you draw a sample of 20 children with this disease and find a mean height of 27.5” and a sample standard deviation of 1.5”.

42. What are the null and alternative hypotheses for this study?

43. What distribution will you use to test your hypothesis, and why?

44. What is the test statistic and the *p*-value?

45. Based on your sample results, what is your decision?

46. Suppose the mean for your sample was 25.0. Redo the calculations and describe what your decision would be.

9.5: Additional Information and Full Hypothesis Test Examples

47. You conduct a study using $\alpha = 0.05$. What is the level of significance for this study?

48. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu = 35.5$
H_a: $\mu \neq 35.5$
Will you conduct a one-tailed or two-tailed test?

49. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu \geq 35.5$
H_a: $\mu < 35.5$
Will you conduct a one-tailed or two-tailed test?

Use the following information to answer the next three exercises. Nationally, 80 percent of adults own an automobile. You are interested in whether the same proportion in your community own cars. You draw a sample of 100 and find that 75 percent own cars.

50. What are the null and alternative hypotheses for this study?

51. What test will you use, and why?

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. You conduct a poll of political opinions, interviewing both members of 50 married couples. Are the groups in this study independent or matched?

53. You are testing a new drug to treat insomnia. You randomly assign 80 volunteer subjects to either the experimental (new drug) or control (standard treatment) conditions. Are the groups in this study independent or matched?

54. You are investigating the effectiveness of a new math textbook for high school students. You administer a pretest to a group of students at the beginning of the semester, and a posttest at the end of a year’s instruction using this textbook, and compare the results. Are the groups in this study independent or matched?

Use the following information to answer the next two exercises. You are conducting a study of the difference in time at two colleges for undergraduate degree completion. At College A, students take an average of 4.8 years to complete an undergraduate degree, while at College B, they take an average of 4.2 years. The pooled standard deviation for this data is 1.6 years

55. Calculate Cohen's *d* and interpret it.

56. Suppose the mean time to earn an undergraduate degree at College A was 5.2 years. Calculate the effect size and interpret it.

57. You conduct an independent-samples t-test with sample size ten in each of two groups. If you are conducting a two-tailed hypothesis test with $\alpha = 0.01$, what p-values will cause you to reject the null hypothesis?

58. You conduct an independent samples *t*-test with sample size 15 in each group, with the following hypotheses:
H_0: $\mu \geq 110$
H_a: $\mu < 110$
If $\alpha = 0.05$, what *t*-values will cause you to reject the null hypothesis?

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

Use the following information to answer the next six exercises. College students in the sciences often complain that they must spend more on textbooks each semester than students in the humanities. To test this, you draw random samples of 50 science and 50 humanities students from your college, and record how much each spent last semester on textbooks. Consider the science students to be group one, and the humanities students to be group two.

59. What is the random variable for this study?

60. What are the null and alternative hypotheses for this study?

61. If the 50 science students spent an average of $530 with a sample standard deviation of $20 and the 50 humanities students spent an average of $380 with a sample standard deviation of $15, would you not reject or reject the null hypothesis? Use an alpha level of 0.05. What is your conclusion?

62. What would be your decision, if you were using $\alpha = 0.01$?

10.3: Comparing Two Independent Population Proportions

Use the information to answer the next six exercises. You want to know if proportion of homes with cable television service differs between Community A and Community B. To test this, you draw a random sample of 100 for each and record whether they have cable service.

63. What are the null and alternative hypotheses for this study

64. If 65 households in Community A have cable service, and 78 households in community B, what is the pooled proportion?

65. At $\alpha = 0.03$, will you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

66. Using an alpha value of 0.01, would you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

10.4: Matched or Paired Samples

Use the following information to answer the next five exercises. You are interested in whether a particular exercise program helps people lose weight. You conduct a study in which you weigh the participants at the start of the study, and again at the conclusion, after they have participated in the exercise program for six months. You

compare the results using a matched-pairs t-test, in which the data is {weight at conclusion – weight at start}. You believe that, on average, the participants will have lost weight after six months on the exercise program.

67. What are the null and alternative hypotheses for this study?

68. Calculate the test statistic, assuming that x_d = –5, s_d = 6, and n = 30 (pairs).

69. What are the degrees of freedom for this statistic?

70. Using α = 0.05, what is your decision regarding the effectiveness of this program in causing weight loss? What is the conclusion?

71. What would it mean if the *t*-statistic had been 4.56, and what would have been your decision in that case?

11.1: Facts About the Chi-Square Distribution

72. What is the mean and standard deviation for a chi-square distribution with 20 degrees of freedom?

11.2: Goodness-of-Fit Test

Use the following information to answer the next four exercises. Nationally, about 66 percent of high school graduates enroll in higher education. You perform a chi-square goodness of fit test to see if this same proportion applies to your high school's most recent graduating class of 200. Your null hypothesis is that the national distribution also applies to your high school.

73. What are the expected numbers of students from your high school graduating class enrolled and not enrolled in higher education?

74. Fill out the rest of this table.

	Observed (*O*)	**Expected (*E*)**	***O* – *E***	**(*O* – *E*)2**	$\frac{(O-E)^2}{z}$
Enrolled	145				
Not enrolled	55				

75. What are the degrees of freedom for this chi-square test?

76. What is the chi-square test statistic and the *p*-value. At the 5% significance level, what do you conclude?

77. For a chi-square distribution with 92 degrees of freedom, the curve ______________.

78. For a chi-square distribution with five degrees of freedom, the curve is _______________.

11.3: Test of Independence

Use the following information to answer the next four exercises. You are considering conducting a chi-square test of independence for the data in this table, which displays data about cell phone ownership for freshman and seniors at a high school. Your null hypothesis is that cell phone ownership is independent of class standing.

79. Compute the expected values for the cells.

	Cell = Yes	**Cell = No**
Freshman	100	150
Senior	200	50

80. Compute $\frac{(O-E)^2}{z}$ for each cell, where O = observed and E = expected.

81. What is the chi-square statistic and degrees of freedom for this study?

82. At the α = 0.5 significance level, what is your decision regarding the null hypothesis?

11.4: Test of Homogeneity

83. You conduct a chi-square test of homogeneity for data in a five by two table. What is the degrees of freedom for this test?

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. A 2013 poll in the State of California surveyed people about taxing sugar-sweetened beverages. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a hypothesis test at the 5% significance level.

Ethnic Group \ Response Type	**Favor**	**Oppose**	**No Opinion**	**Row Total**
White / Non-Hispanic	234	433	43	710
Latino	147	106	19	272
African American	24	41	6	71
Asian American	54	48	16	118
Column Total	459	628	84	1171

85. In a test of homogeneity, what must be true about the expected value of each cell?

86. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of independence?

87. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of homogeneity?

11.6: Test of a Single Variance

88. A lab test claims to have a variance of no more than five. You believe the variance is greater. What are the null and alternative hypothesis to test this?

Practice Test 3 Solutions

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

1. $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

2. normal

3. 0.025 or 2.5%; A 95% confidence interval contains 95% of the probability, and excludes five percent, and the five percent excluded is split evenly between the upper and lower tails of the distribution.

4. *z*-score = 1.96; $EBM = \ z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)\,(0.73) = \ 1.4308$

5. 41 ± 1.43 = (39.57, 42.43); Using the calculator function Zinterval, answer is (40.74, 41.26. Answers differ due to rounding.

6. The *z*-value for a 90% confidence interval is 1.645, so *EBM* = 1.645(0.73) = 1.20085.
The 90% confidence interval is 41 ± 1.20 = (39.80, 42.20).
The calculator function Zinterval answer is (40.78, 41.23). Answers differ due to rounding.

7. The standard error of measurement is: $\frac{\sigma}{\sqrt{n}} = \ \frac{4}{\sqrt{50}} = 0.57$
$EBM = \ z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)\,(0.57) = \ 1.12$
The 95% confidence interval is 41 ± 1.12 = (39.88, 42.12).
The calculator function Zinterval answer is (40.84, 41.16). Answers differ due to rounding.

8. The 99% confidence interval, because it includes all but one percent of the distribution. The 95% confidence interval will be narrower, because it excludes five percent of the distribution.

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's t

9. The *t*-distribution will have more probability in its tails ("thicker tails") and less probability near the mean of the distribution ("shorter in the center").

10. Both distributions are symmetrical and centered at zero.

11. *df* = *n* – 1 = 20 – 1 = 19

12. You can get the *t*-value from a probability table or a calculator. In this case, for a *t*-distribution with 19 degrees of freedom, and a 95% two-sided confidence interval, the value is 2.093, i.e.,
$t_{\frac{\alpha}{2}} \ = \mathbf{2.093}$. The calculator function is invT(0.975, 19).

13. $EBM = \ t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.093)\left(\frac{0.3}{\sqrt{20}}\right) = \ 0.140$
98.4 ± 0.14 = (98.26, 98.54).
The calculator function Tinterval answer is (98.26, 98.54).

14. $t_{\frac{\alpha}{2}} = 2.861$. The calculator function is invT(0.995, 19).

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.861)\left(\frac{0.3}{\sqrt{20}}\right) = 0.192$$

98.4 ± 0.19 = (98.21, 98.59). The calculator function Tinterval answer is (98.21, 98.59).

15. *df* = *n* – 1 = 30 – 1 = 29. $t_{\frac{\alpha}{2}} \ = 2.045$
$EBM = \ z_t\left(\frac{s}{\sqrt{n}}\right) = (2.045)\left(\frac{0.3}{\sqrt{30}}\right) = \ 0.112$
98.4 ± 0.11 = (98.29, 98.51). The calculator function Tinterval answer is (98.29, 98.51).

8.3: Confidence Interval for a Population Proportion

16. $p' = \frac{280}{500} = 0.56$
$q' = 1 - p' = 1 - 0.56 = 0.44$
$s = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.56(0.44)}{500}} = 0.0222$

17. Because you are using the normal approximation to the binomial, $z_{\frac{\alpha}{2}} = 1.96$.
Calculate the error bound for the population (*EBP*):
$EBP = \ z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.96\,(0.222) = 0.0435$
Calculate the 95% confidence interval:
0.56 ± 0.0435 = (0.5165, 0.6035).
The calculator function 1-PropZint answer is (0.5165, 0.6035).

18. $z_{\frac{\alpha}{2}} = 1.64$
$EBP = \ z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.64\,(0.0222) = 0.0364$
0.56 ± 0.03 = (0.5236, 0.5964). The calculator function 1-PropZint answer is (0.5235, 0.5965)

19. $z_{\frac{\alpha}{2}} = 2.58$
$EBP = \ z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 2.58\,(0.0222) = \ 0.0573$
0.56 ± 0.05 = (0.5127, 0.6173).
The calculator function 1-PropZint answer is (0.5028, 0.6172).

20. *EBP* = 0.04 (because 4% = 0.04)
$z_{\frac{\alpha}{2}} = 1.96$ for a 95% confidence interval
$n = \ \frac{z^2pq}{EBP^2} = \ \frac{1.96^2(0.5)(0.5)}{0.04^2} = \ \frac{0.9604}{0.0016} = 600.25$
You need 601 subjects (rounding upward from 600.25).

21. $n = \ \frac{n^2pq}{EBP^2} = \ \frac{1.96^2(0.6)(0.4)}{0.04^2} = \ \frac{0.9220}{0.0016} = 576.24$
You need 577 subjects (rounding upward from 576.24).

22. $n = \ \frac{n^2pq}{EBP^2} = \ \frac{1.96^2(0.5)(0.5)}{0.03^2} = \ \frac{0.9604}{0.0009} = 1067.11$
You need 1,068 subjects (rounding upward from 1,067.11).

9.1: Null and Alternate Hypotheses

23. H_0: $p = 0.58$
H_a: $p \neq 0.58$

24. H_0: $p \geq 0.58$
H_a: $p < 0.58$

25. H_0: $\mu \geq \$268,000$
H_a: $\mu < \$268,000$

26. H_a: $\mu \neq 107$

27. H_a: $p \geq 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. a Type I error

29. a Type II error

30. Power = $1 - \beta = 1 - P$(Type II error).

31. The null hypothesis is that the patient does not have cancer. A Type I error would be detecting cancer when it is not present. A Type II error would be not detecting cancer when it is present. A Type II error is more serious, because failure to detect cancer could keep a patient from receiving appropriate treatment.

32. The screening test has a ten percent probability of a Type I error, meaning that ten percent of the time, it will detect TB when it is not present.

33. The screening test has a 20 percent probability of a Type II error, meaning that 20 percent of the time, it will fail to detect TB when it is in fact present.

34. Eighty percent of the time, the screening test will detect TB when it is actually present.

9.3: Distribution Needed for Hypothesis Testing

35. The Student's *t*-test.

36. The normal distribution or *z*-test.

37. The normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$

38. t_{24}. You use the *t*-distribution because you don't know the population standard deviation, and the degrees of freedom are 24 because $df = n - 1$.

39. $X \sim N\left(0.95, \frac{0.051}{\sqrt{100}}\right)$
Because you know the population standard deviation, and have a large sample, you can use the normal distribution.

9.4: Rare Events, the Sample, Decision, and Conclusion

40. Fail to reject the null hypothesis, because $\alpha \leq p$

41. Reject the null hypothesis, because $\alpha \geq p$.

42. H_0: $\mu \geq 29.0$”
H_a: $\mu < 29.0$”

43. t_{19}. Because you do not know the population standard deviation, use the *t*-distribution. The degrees of freedom are 19, because $df = n - 1$.

44. The test statistic is –4.4721 and the *p*-value is 0.00013 using the calculator function TTEST.

45. With $\alpha = 0.05$, reject the null hypothesis.

46. With $\alpha = 0.05$, the *p*-value is almost zero using the calculator function TTEST so reject the null hypothesis.

9.5: Additional Information and Full Hypothesis Test Examples

47. The level of significance is five percent.

48. two-tailed

49. one-tailed

50. H_0: $p = 0.8$
H_a: $p \neq 0.8$

51. You will use the normal test for a single population proportion because *np* and *nq* are both greater than five.

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. They are matched (paired), because you interviewed married couples.

53. They are independent, because participants were assigned at random to the groups.

54. They are matched (paired), because you collected data twice from each individual.

55. $d = \frac{x_1 - x_2}{s_{pooled}} = \frac{4.8 - 4.2}{1.6} = 0.375$
This is a small effect size, because 0.375 falls between Cohen’s small (0.2) and medium (0.5) effect sizes.

56. $d = \frac{x_1 - x_2}{s_{pooled}} = \frac{5.2 - 4.2}{1.6} = 0.625$
The effect size is 0.625. By Cohen’s standard, this is a medium effect size, because it falls between the medium (0.5) and large (0.8) effect sizes.

57. *p*-value < 0.01.

58. You will only reject the null hypothesis if you get a value significantly below the hypothesized mean of 110.

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

59. $X_1 - X_2$, i.e., the mean difference in amount spent on textbooks for the two groups.

60. H_0: $X_1 - X_2 \leq 0$
H_a: $X_1 - X_2 > 0$
This could also be written as:
H_0: $X_1 \leq X_2$
H_a: $X_1 > X_2$

61. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

62. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 1% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

10.3: Comparing Two Independent Population Proportions

63. H_0: $p_A = p_B$
H_a: $p_A \neq p_B$

64. $p_c = \frac{x_A + x_A}{n_A + n_A} = \frac{65+78}{100+100} = 0.715$

65. Using the calculator function 2-PropZTest, the p-value = 0.0417. Reject the null hypothesis. At the 3% significance level, here is sufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

66. Using the calculator function 2-PropZTest, the *p*-value = 0.0417. Do not reject the null hypothesis. At the 1% significance level, there is insufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

10.4: Matched or Paired Samples

67. H_0: $x_d \geq 0$
H_a: $x_d < 0$

68. t = – 4.5644

69. df = 30 – 1 = 29.

70. Using the calculator function TTEST, the *p*-value = 0.00004 so reject the null hypothesis. At the 5% level, there is sufficient evidence to conclude that the participants lost weight, on average.

71. A positive *t*-statistic would mean that participants, on average, gained weight over the six months.

11.1: Facts About the Chi-Square Distribution

72. $\mu = df = 20$
$\sigma = \sqrt{2(df)} = \sqrt{40} = 6.32$

11.2: Goodness-of-Fit Test

73. Enrolled = 200(0.66) = 132. Not enrolled = 200(0.34) = 68

74.

	Observed (O)	**Expected (E)**	**O – E**	**(O – E)2**	$\frac{(O-E)^2}{z}$
Enrolled	145	132	145 – 132 = 13	169	$\frac{169}{132} = 1.280$
Not enrolled	55	68	55 – 68 = –13	169	$\frac{169}{68} = 2.485$

75. $df = n - 1 = 2 - 1 = 1$.

76. Using the calculator function Chi-square GOF – Test (in STAT TESTS), the test statistic is 3.7656 and the p-value is 0.0523. Do not reject the null hypothesis. At the 5% significance level, there is insufficient evidence to conclude that high school most recent graduating class distribution of enrolled and not enrolled does not fit that of the national distribution.

77. approximates the normal

78. skewed right

11.3: Test of Independence

79.

	Cell = Yes	**Cell = No**	**Total**
Freshman	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Senior	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Total	300	200	500

80. $\frac{(100-150)^2}{150} = 16.67$
$\frac{(150-100)^2}{100} = 25$
$\frac{(200-100)^2}{150} = 16.67$
$\frac{(50-100)^2}{100} = 25$

81. Chi-square = 16.67 + 25 + 16.67 + 25 = 83.34.
$df = (r - 1)(c - 1) = 1$

82. p-value = P(Chi-square, 83.34) = 0
Reject the null hypothesis.
You could also use the calculator function STAT TESTS Chi-Square – Test.

11.4: Test of Homogeneity

83. The table has five rows and two columns. $df = (r – 1)(c – 1) = (4)(1) = 4$.

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. Using the calculator function (STAT TESTS) Chi-square Test, the p-value = 0. Reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the poll responses independent of the participants' ethnic group.

85. The expected value of each cell must be at least five.

86. H_0: The variables are independent.
H_a: The variables are not independent.

87. H_0: The populations have the same distribution.
H_a: The populations do not have the same distribution.

11.6: Test of a Single Variance

88. H_0: $\sigma^2 \leq 5$
H_a: $\sigma^2 > 5$

Practice Test 4

12.1 Linear Equations

1. Which of the following equations is/are linear?

a. $y = –3x$
b. $y = 0.2 + 0.74x$
c. $y = –9.4 – 2x$
d. A and B
e. A, B, and C

2. To complete a painting job requires four hours setup time plus one hour per 1,000 square feet. How would you express this information in a linear equation?

3. A statistics instructor is paid a per-class fee of $2,000 plus $100 for each student in the class. How would you express this information in a linear equation?

4. A tutoring school requires students to pay a one-time enrollment fee of $500 plus tuition of $3,000 per year. Express this information in an equation.

12.2: Slope and *Y*-intercept of a Linear Equation

Use the following information to answer the next four exercises. For the labor costs of doing repairs, an auto mechanic charges a flat fee of $75 per car, plus an hourly rate of $55.

5. What are the independent and dependent variables for this situation?

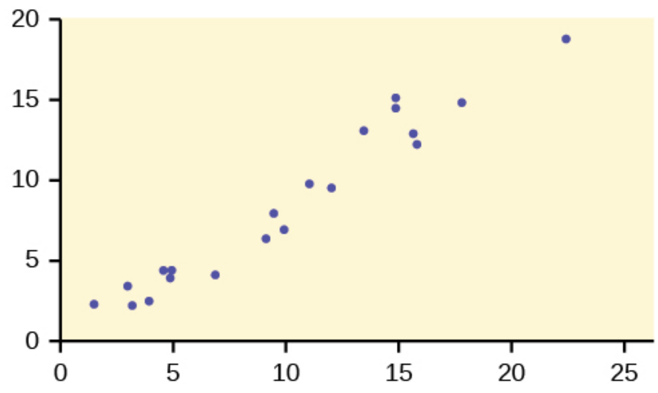
6. Write the equation and identify the slope and intercept.

7. What is the labor charge for a job that takes 3.5 hours to complete?

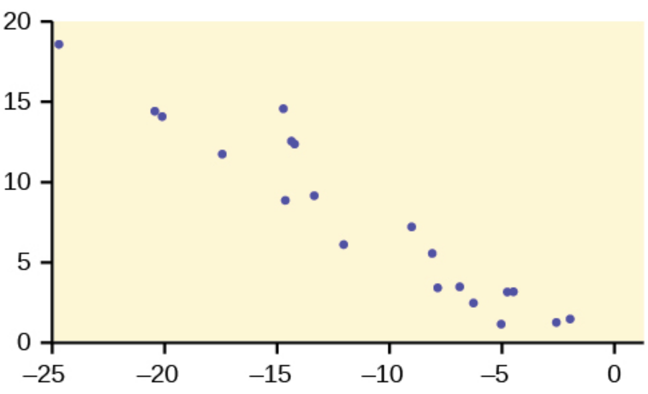
8. One job takes 2.4 hours to complete, while another takes 6.3 hours. What is the difference in labor costs for these two jobs?

12.3: Scatter Plots

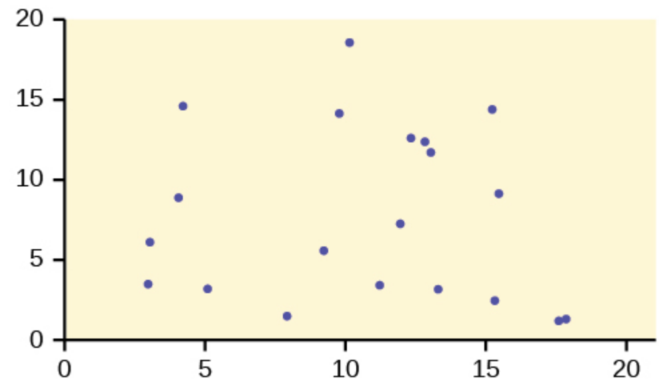
9. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



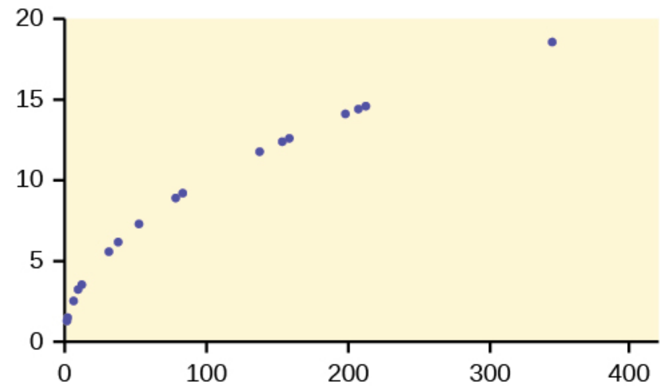
10. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



11. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



12. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



12.4: The Regression Equation

Use the following information to answer the next four exercises. Height (in inches) and weight (In pounds) in a sample of college freshman men have a linear relationship with the following summary statistics:
x = 68.4
y =141.6
s_x = 4.0
s_y = 9.6
r = 0.73
Let Y = weight and X = height, and write the regression equation in the form:
$\hat{y} = a + bx$

13. What is the value of the slope?

14. What is the value of the *y* intercept?

15. Write the regression equation predicting weight from height in this data set, and calculate the predicted weight for someone 68 inches tall.

12.5: Correlation Coefficient and Coefficient of Determination

16. The correlation between body weight and fuel efficiency (measured as miles per gallon) for a sample of 2,012 model cars is –0.56. Calculate the coefficient of determination for this data and explain what it means.

17. The correlation between high school GPA and freshman college GPA for a sample of 200 university students is 0.32. How much variation in freshman college GPA is not explained by high school GPA?

18. Rounded to two decimal places what correlation between two variables is necessary to have a coefficient of determination of at least 0.50?

12.6: Testing the Significance of the Correlation Coefficient

19. Write the null and alternative hypotheses for a study to determine if two variables are significantly correlated.

20. In a sample of 30 cases, two variables have a correlation of 0.33. Do a *t*-test to see if this result is significant at the $\alpha = 0.05$ level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

21. In a sample of 25 cases, two variables have a correlation of 0.45. Do a *t*-test to see if this result is significant at the $\alpha = 0.05$ level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

12.7: Prediction

Use the following information to answer the next two exercises. A study relating the grams of potassium (Y) to the grams of fiber (X) per serving in enriched flour products (bread, rolls, etc.) produced the equation:

$$\hat{y} = 25 + 16x$$

22. For a product with five grams of fiber per serving, what are the expected grams of potassium per serving?

23. Comparing two products, one with three grams of fiber per serving and one with six grams of fiber per serving, what is the expected difference in grams of potassium per serving?

12.8: Outliers

24. In the context of regression analysis, what is the definition of an outlier, and what is a rule of thumb to evaluate if a given value in a data set is an outlier?

25. In the context of regression analysis, what is the definition of an influential point, and how does an influential point differ from an outlier?

26. The least squares regression line for a data set is $\hat{y} = 5 + 0.3x$ and the standard deviation of the residuals is 0.4. Does a case with the values $x = 2$, $y = 6.2$ qualify as an outlier?

27. The least squares regression line for a data set is $\hat{y} = 2.3 - 0.1x$ and the standard deviation of the residuals is 0.13. Does a case with the values $x = 4.1$, $y = 2.34$ qualify as an outlier?

13.1: One-Way ANOVA

28. What are the five basic assumptions to be met if you want to do a one-way ANOVA?

29. You are conducting a one-way ANOVA comparing the effectiveness of four drugs in lowering blood pressure in hypertensive patients. What are the null and alternative hypotheses for this study?

30. What is the primary difference between the independent samples *t*-test and one-way ANOVA?

31. You are comparing the results of three methods of teaching geometry to high school students. The final exam scores X_1, X_2, X_3, for the samples taught by the different methods have the following distributions:
$X_1 \sim N(85, 3.6)$
$X_1 \sim N(82, 4.8)$
$X_1 \sim N(79, 2.9)$
Each sample includes 100 students, and the final exam scores have a range of 0–100. Assuming the samples are independent and randomly selected, have the requirements for conducting a one-way ANOVA been met? Explain why or why not for each assumption.

32. You conduct a study comparing the effectiveness of four types of fertilizer to increase crop yield on wheat farms. When examining the sample results, you find that two of the samples have an approximately normal distribution, and two have an approximately uniform distribution. Is this a violation of the assumptions for conducting a one-way ANOVA?

13.2: The *F* Distribution

Use the following information to answer the next seven exercises. You are conducting a study of three types of feed supplements for cattle to test their effectiveness in producing weight gain among calves whose feed includes one of the supplements. You have four groups of 30 calves (one is a control group receiving the usual feed, but no supplement). You will conduct a one-way ANOVA after one year to see if there are difference in the mean weight for the four groups.

33. What is SS_{within} in this experiment, and what does it mean?

34. What is $SS_{between}$ in this experiment, and what does it mean?

35. What are *k* and *i* for this experiment?

36. If $SS_{within} = 374.5$ and $SS_{total} = 621.4$ for this data, what is $SS_{between}$?

37. What are $MS_{between}$, and MS_{within}, for this experiment?

38. What is the *F* Statistic for this data?

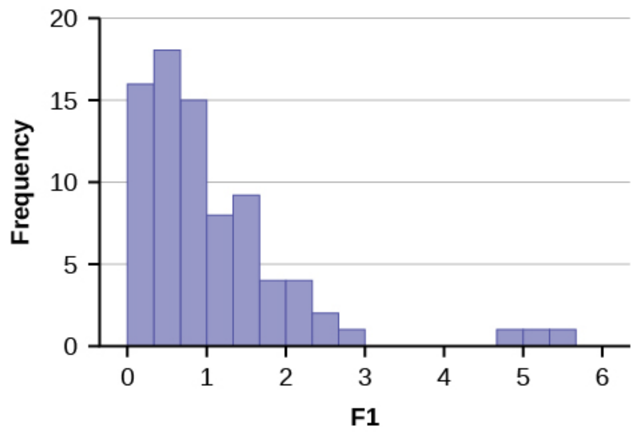
39. If there had been 35 calves in each group, instead of 30, with the sums of squares remaining the same, would the *F* Statistic be larger or smaller?

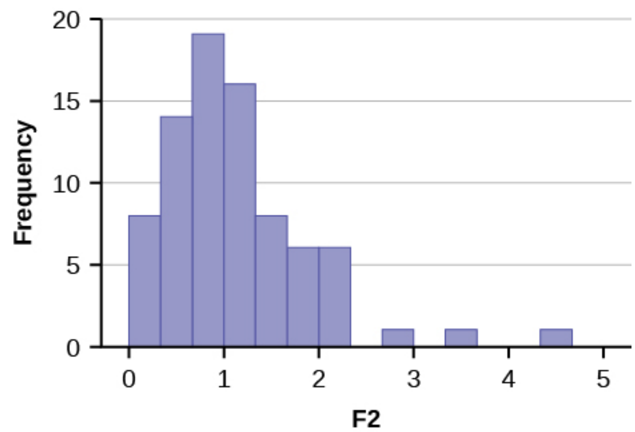
13.3: Facts About the *F* Distribution

40. Which of the following numbers are possible *F* Statistics?

a. 2.47
b. 5.95
c. –3.61
d. 7.28
e. 0.97

41. Histograms *F*1 and *F*2 below display the distribution of cases from samples from two populations, one distributed $F_{3,15}$ and one distributed $F_{5,500}$. Which sample came from which population?





42. The F Statistic from an experiment with $k = 3$ and $n = 50$ is 3.67. At $\alpha = 0.05$, will you reject the null hypothesis?

43. The F Statistic from an experiment with $k = 4$ and $n = 100$ is 4.72. At $\alpha = 0.01$, will you reject the null hypothesis?

13.4: Test of Two Variances

44. What assumptions must be met to perform the F test of two variances?

45. You believe there is greater variance in grades given by the math department at your university than in the English department. You collect all the grades for undergraduate classes in the two departments for a semester, and compute the variance of each, and conduct an F test of two variances. What are the null and alternative hypotheses for this study?

Practice Test 4 Solutions

12.1 Linear Equations

1. e. A, B, and C.
All three are linear equations of the form $y = mx + b$.

2. Let y = the total number of hours required, and x the square footage, measured in units of 1,000. The equation is: $y = x + 4$

3. Let y = the total payment, and x the number of students in a class. The equation is: $y = 100(x) + 2,000$

4. Let y = the total cost of attendance, and x the number of years enrolled. The equation is: $y = 3,000(x) + 500$

12.2: Slope and Y-intercept of a Linear Equation

5. The independent variable is the hours worked on a car. The dependent variable is the total labor charges to fix a car.

6. Let y = the total charge, and x the number of hours required. The equation is: $y = 55x + 75$
The slope is 55 and the intercept is 75.

7. $y = 55(3.5) + 75 = 267.50$

8. Because the intercept is included in both equations, while you are only interested in the difference in costs, you do not need to include the intercept in the solution. The difference in number of hours required is: 6.3 – 2.4 = 3.9. Multiply this difference by the cost per hour: 55(3.9) = 214.5.
The difference in cost between the two jobs is $214.50.

12.3: Scatter Plots

9. The X and Y variables have a strong linear relationship. These variables would be good candidates for analysis with linear regression.

10. The X and Y variables have a strong negative linear relationship. These variables would be good candidates for analysis with linear regression.

11. There is no clear linear relationship between the X and Y variables, so they are not good candidates for linear regression.

12. The X and Y variables have a strong positive relationship, but it is curvilinear rather than linear. These variables are not good candidates for linear regression.

12.4: The Regression Equation

13. $r\left(\frac{s_y}{s_x}\right) = 0.73\left(\frac{9.6}{4.0}\right) = 1.752 \approx 1.75$

14. $a = y - bx = 141.6 - 1.752(68.4) = 21.7632 \approx 21.76$

15. $\hat{y} = 21.76 + 1.75(68) = 140.76$

12.5: Correlation Coefficient and Coefficient of Determination

16. The coefficient of determination is the square of the correlation, or r^2.
For this data, $r^2 = (–0.56)2 = 0.3136 \approx 0.31$ or 31%. This means that 31 percent of the variation in fuel efficiency can be explained by the bodyweight of the automobile.

17. The coefficient of determination = 0.32^2 = 0.1024. This is the amount of variation in freshman college GPA that can be explained by high school GPA. The amount that cannot be explained is 1 – 0.1024 = 0.8976 ≈ 0.90. So about 90 percent of variance in freshman college GPA in this data is not explained by high school GPA.

18. $r = \sqrt{r^2}$
$\sqrt{0.5} = 0.707106781 \approx 0.71$
You need a correlation of 0.71 or higher to have a coefficient of determination of at least 0.5.

12.6: Testing the Significance of the Correlation Coefficient

19. H_0: $\rho = 0$
H_a: $\rho \neq 0$

20. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.33\sqrt{30-2}}{\sqrt{1-0.33^2}} = 1.85$
The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{29} distribution is 2.045. Your value is less than this, so you fail to reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.
Using the calculator function tcdf, the *p*-value is 2tcdf(1.85, 10^99, 29) = 0.0373. Do not reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

21. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{25-2}}{\sqrt{1-0.45^2}} = 2.417$
The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{24} distribution is 2.064. Your value is greater than this, so you reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.
Using the calculator function tcdf, the p-value is 2tcdf(2.417, 10^99, 24) = 0.0118. Reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

12.7: Prediction

22. $\hat{y} = 25 + 16(5) = 105$

23. Because the intercept appears in both predicted values, you can ignore it in calculating a predicted difference score. The difference in grams of fiber per serving is 6 – 3 = 3 and the predicted difference in grams of potassium per serving is (16)(3) = 48.

12.8: Outliers

24. An outlier is an observed value that is far from the least squares regression line. A rule of thumb is that a point more than two standard deviations of the residuals from its predicted value on the least squares regression line is an outlier.

25. An influential point is an observed value in a data set that is far from other points in the data set, in a horizontal direction. Unlike an outlier, an influential point is determined by its relationship with other values in the data set, not by its relationship to the regression line.

26. The predicted value for y is: $\hat{y} = 5 + 0.3x = 5.6$. The value of 6.2 is less than two standard deviations from the predicted value, so it does not qualify as an outlier.
Residual for (2, 6.2): 6.2 – 5.6 = 0.6 (0.6 < 2(0.4))

27. The predicted value for y is: $\hat{y}$ = 2.3 – 0.1(4.1) = 1.89. The value of 2.32 is more than two standard deviations from the predicted value, so it qualifies as an outlier.
Residual for (4.1, 2.34): 2.32 – 1.89 = 0.43 (0.43 > 2(0.13))

13.1: One-Way ANOVA

28.

1. Each sample is drawn from a normally distributed population
2. All samples are independent and randomly selected.
3. The populations from which the samples are draw have equal standard deviations.
4. The factor is a categorical variable.
5. The response is a numerical variable.

29. H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$
H_a: At least two of the group means $\mu 1$, $\mu 2$, $\mu 3$, $\mu 4$ are not equal.

30. The independent samples *t*-test can only compare means from two groups, while one-way ANOVA can compare means of more than two groups.

31. Each sample appears to have been drawn from a normally distributed populations, the factor is a categorical variable (method), the outcome is a numerical variable (test score), and you were told the samples were independent and randomly selected, so those requirements are met. However, each sample has a different standard deviation, and this suggests that the populations from which they were drawn also have different standard deviations, which is a violation of an assumption for one-way ANOVA. Further statistical testing will be necessary to test the assumption of equal variance before proceeding with the analysis.

32. One of the assumptions for a one-way ANOVA is that the samples are drawn from normally distributed populations. Since two of your samples have an approximately uniform distribution, this casts doubt on whether this assumption has been met. Further statistical testing will be necessary to determine if you can proceed with the analysis.

13.2: The *F* Distribution

33. SS_{within} is the sum of squares within groups, representing the variation in outcome that cannot be attributed to the different feed supplements, but due to individual or chance factors among the calves in each group.

34. $SS_{between}$ is the sum of squares between groups, representing the variation in outcome that can be attributed to the different feed supplements.

35. k = the number of groups = 4
n_1 = the number of cases in group 1 = 30
n = the total number of cases = 4(30) = 120

36. $SS_{total} = SS_{within} + SS_{between}$ so $SS_{between} = SS_{total} - SS_{within}$
621.4 – 374.5 = 246.9

37. The mean squares in an ANOVA are found by dividing each sum of squares by its respective degrees of freedom (*df*).
For SS_{total}, $df = n - 1 = 120 - 1 = 119$.
For $SS_{between}$, df = k – 1 = 4 – 1 = 3.
For SS_{within}, df = 120 – 4 = 116.
$MS_{between} = \frac{246.9}{3} = 82.3$
$MS_{within} = \frac{374.5}{116} = 3.23$

38. $F = \frac{MS_{between}}{MS_{within}} = \frac{82.3}{3.23} = 25.48$

39. It would be larger, because you would be dividing by a smaller number. The value of $MS_{between}$ would not change with a change of sample size, but the value of MS_{within} would be smaller, because you would be dividing by a larger number (df_{within} would be 136, not 116). Dividing a constant by a smaller number produces a larger result.

13.3: Facts About the *F* Distribution

40. All but choice c, –3.61. *F* Statistics are always greater than or equal to 0.

41. As the degrees of freedom increase in an *F* distribution, the distribution becomes more nearly normal. Histogram *F*2 is closer to a normal distribution than histogram *F*1, so the sample displayed in histogram *F*1 was drawn from the $F_{3,15}$ population, and the sample displayed in histogram *F*2 was drawn from the $F_{5,500}$ population.

42. Using the calculator function Fcdf, *p*-value = Fcdf(3.67, 1E, 3,50) = 0.0182. Reject the null hypothesis.

43. Using the calculator function Fcdf, *p*-value = Fcdf(4.72, 1E, 4, 100) = 0.0016 Reject the null hypothesis.

13.4: Test of Two Variances

44. The samples must be drawn from populations that are normally distributed, and must be drawn from independent populations.

45. Let σ_M^2 = variance in math grades, and σ_E^2 = variance in English grades.
H_0: $\sigma_M^2 \leq \sigma_E^2$
H_a: $\sigma_M^2 > \sigma_E^2$

Practice Final Exam 1

Use the following information to answer the next two exercises: An experiment consists of tossing two, 12-sided dice (the numbers 1–12 are printed on the sides of each die).

- Let Event *A* = both dice show an even number.
- Let Event *B* = both dice show a number more than eight

1. Events *A* and *B* are:

a. mutually exclusive.
b. independent.
c. mutually exclusive and independent.
d. neither mutually exclusive nor independent.

2. Find *P*(*A*|*B*).

a. $\frac{2}{4}$
b. $\frac{16}{144}$
c. $\frac{4}{16}$
d. $\frac{2}{144}$

3. Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

a. Sample sizes are almost never small.
b. Two measurements are drawn from the same pair of individuals or objects.

c. Two sample means are compared to each other.
d. Answer choices b and c are both true.

Use the following information to answer the next two exercises: One hundred eighteen students were asked what type of color their bedrooms were painted: light colors, dark colors, or vibrant colors. The results were tabulated according to gender.

	Light colors	**Dark colors**	**Vibrant colors**
Female	20	22	28
Male	10	30	8

4. Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.

a. $\frac{10}{118}$
b. $\frac{68}{118}$
c. $\frac{48}{118}$
d. $\frac{10}{48}$

5. Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.

a. $\frac{30}{118}$
b. $\frac{30}{48}$
c. $\frac{22}{118}$
d. $\frac{30}{52}$

Use the following information to answer the next two exercises: We are interested in the number of times a teenager must be reminded to do his or her chores each week. A survey of 40 mothers was conducted. [link] shows the results of the survey.

x	***P* (*x*)**
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$

x	$P(x)$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

6. Find the probability that a teenager is reminded two times.

a. 8
b. $\frac{8}{40}$
c. $\frac{6}{40}$
d. 2

7. Find the expected number of times a teenager is reminded to do his or her chores.

a. 15
b. 2.78
c. 1.0
d. 3.13

Use the following information to answer the next two exercises: On any given day, approximately 37.5% of the cars parked in the De Anza parking garage are parked crookedly. We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

8. For every 22 cars, how many would you expect to be parked crookedly, on average?

a. 8.25
b. 11
c. 18
d. 7.5

9. What is the probability that at least ten of the 22 cars are parked crookedly.

a. 0.1263
b. 0.1607
c. 0.2870
d. 0.8393

10. Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ scores is 15 points. The correct distribution to use for the hypothesis test is:

a. Binomial
b. Student's *t*
c. Normal
d. Uniform

Use the following information to answer the next three exercises: De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1,795 students enrolled in Math 1A (1st quarter calculus), 1,428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

11. If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be:

a. H_a: $p_A = p_B$
b. H_a: $p_A > p_B$
c. H_o: $p_A = p_B$
d. H_a: $p_A \neq p_B$

12. The Type I error is to:

a. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.
b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.
c. conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.
d. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

13. The correct decision is to:

a. reject H_0
b. not reject H_0
c. There is not enough information given to conduct the hypothesis test

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	**School Average Running Time**	**School Standard Deviation**
Kia	4.9	5.2	0.15
Alejandra	4.2	4.6	0.25
Iris	4.5	4.9	0.12

14. Which student is the BEST when compared to the other runners at her school?

a. Kia
b. Alejandra
c. Iris
d. Impossible to determine

Use the following information to answer the next two exercises: The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog: \$212, \$292, \$278, \$199, \$280, \$236

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least \$275.

15. The correct distribution to use for the hypothesis test is:

a. Normal
b. Binomial

c. Student's *t*
d. Exponential

16. The hypothesis test:

a. is two-tailed.
b. is left-tailed.
c. is right-tailed.
d. has no tails.

17. Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected two buildings on campus and asked each professor in the selected buildings how many books are in his or her office. Sara surveyed 25 professors. The type of sampling selected is

a. simple random sampling.
b. systematic sampling.
c. cluster sampling.
d. stratified sampling.

18. A clothing store would use which measure of the center of data when placing orders for the typical "middle" customer?

a. mean
b. median
c. mode
d. IQR

19. In a hypothesis test, the *p*-value is

a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.
b. called the preconceived alpha.
c. compared to beta to decide whether to reject or not reject the null hypothesis.
d. Answer choices A and B are both true.

Use the following information to answer the next three exercises: A community college offers classes 6 days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her 5 classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in [link].

Number of Days on Campus	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
6	1	.02	1.00

20. Combined with convenience sampling, what other sampling technique did Maria use?

a. simple random
b. systematic
c. cluster
d. stratified

21. How many students come to campus for classes four days a week?

a. 49
b. 25
c. 30
d. 13

22. What is the 60th percentile for the this data?

a. 2
b. 3
c. 4
d. 5

Use the following information to answer the next two exercises: The following data are the results of a random survey of 110 Reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

23. Construct a 95% confidence interval for the true population mean number of dependents of Reservists called to active duty to increase security at California airports.

a. (1.85, 2.32)
b. (1.80, 2.36)
c. (1.97, 2.46)
d. (1.92, 2.50)

24. The 95% confidence interval above means:

a. Five percent of confidence intervals constructed this way will not contain the true population aveage number of dependents.
b. We are 95% confident the true population mean number of dependents falls in the interval.
c. Both of the above answer choices are correct.
d. None of the above.

25. $X \sim U(4, 10)$. Find the 30th percentile.

a. 0.3000
b. 3
c. 5.8
d. 6.1

26. If $X \sim Exp(0.8)$, then $P(x < \mu) =$ ___________

a. 0.3679
b. 0.4727
c. 0.6321
d. cannot be determined

27. The lifetime of a computer circuit board is normally distributed with a mean of 2,500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2,560 hours?

a. 0.8413
b. 0.1587
c. 0.3461
d. 0.6539

28. A survey of 123 reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98% confidence interval for the true population proportion of reservists called to active duty that are married.

a. (0.6030, 0.7954)
b. (0.6181, 0.7802)
c. (0.5927, 0.8057)
d. (0.6312, 0.7672)

29. Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last ten marathon winning times is collected. Let x = mean winning times for ten marathons. The distribution for x is:

a. $N\left(145, \frac{14}{\sqrt{10}}\right)$
b. $N\,(145,14)$
c. t_9
d. t_{10}

30. Suppose that Phi Beta Kappa honors the top one percent of college and university seniors. Assume that grade point means (GPA) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum GPA needed to become a member of Phi Beta Kappa at that college?

a. 3.99
b. 1.34
c. 3.00
d. 3.66

The number of people living on American farms has declined steadily during the 20th century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	198
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

31. The linear regression equation is $\hat{y}$ = 1166.93 – 0.5868*x*. What was the expected farm population (in millions of persons) for 1980?

a. 7.2
b. 5.1
c. 6.0
d. 8.0

32. In linear regression, which is the best possible *SSE*?

a. 13.46
b. 18.22
c. 24.05
d. 16.33

33. In regression analysis, if the correlation coefficient is close to one what can be said about the best fit line?

a. It is a horizontal line. Therefore, we can not use it.
b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.
c. The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.
d. We do not have the equation. Therefore, we cannot say anything about it.

Use the following information to answer the next three exercises: A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	**Male**
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Does the data suggest that there is a relationship between the gender of students and their choice of major?

34. The distribution for the test is:

a. $\text{Chi}^2{}_8$.
b. $\text{Chi}^2{}_3$.
c. t_{721}.
d. $N\left(0,1\right)$.

35. The expected number of female who choose finance is:

a. 37.
b. 61.
c. 60.
d. 70.

36. The *p*-value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

a. there is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
b. there is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
c. there is sufficient evidence to conclude that students find economics very hard.
d. there is in sufficient evidence to conclude that more females prefer administration than males.

37. An agency reported that the work force nationwide is composed of 10% professional, 10% clerical, 30% skilled, 15% service, and 35% semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = 0.10$ does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

a. Chi^2 goodness of fit
b. Chi^2 test of independence
c. Independent groups proportions
d. Unable to determine

Practice Final Exam 1 Solutions

Solutions

1. b. independent

2. c. $\frac{4}{16}$

3. b. Two measurements are drawn from the same pair of individuals or objects.

4. b. $\frac{68}{118}$

5. d. $\frac{30}{52}$

6. b. $\frac{8}{40}$

7. b. 2.78

8. a. 8.25

9. c. 0.2870

10. c. Normal

11. d. H_a: $p_A \neq p_B$

12. b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

13. b. not reject H_0

14. c. Iris

15. c. Student's *t*

16. b. is left-tailed.

17. c. cluster sampling

18. b. median

19. a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

20. d. stratified

21. b. 25

22. c. 4

23. a. (1.85, 2.32)

24. c. Both above are correct.

25. c. 5.8

26. c. 0.6321

27. a. 0.8413

28. a. (0.6030, 0.7954)

29. a. $N\left(145, \frac{14}{\sqrt{10}}\right)$

30. d. 3.66

31. b. 5.1

32. a. 13.46

33. b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.

34. b. $\text{Chi}^2{}_3$.

35. d. 70

36. b. There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

37. a. Chi^2 goodness-of-fit

Practice Final Exam 2

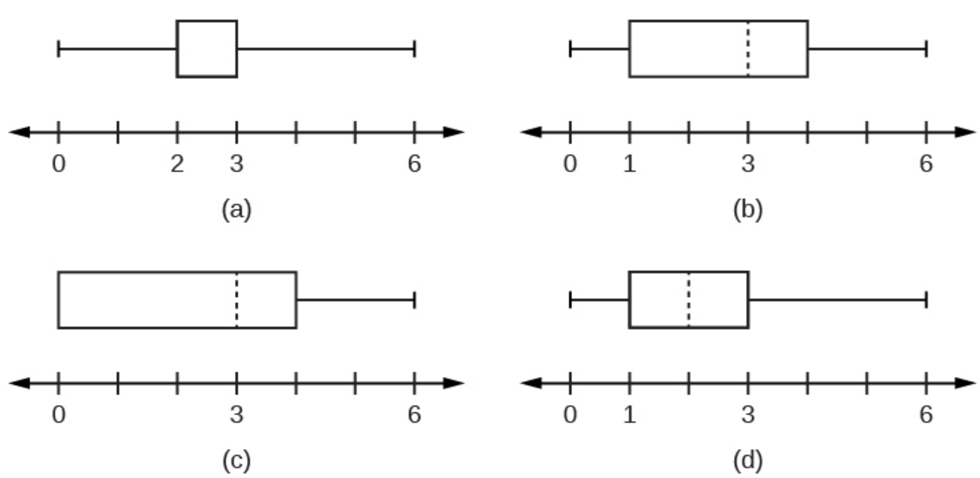
1. A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the:

a. statistic.
b. parameter.
c. population.
d. variable.

Use the following information to answer the next two exercises:

value	**frequency**
0	1
1	4
2	7
3	9
6	4

2. The box plot for the data is:



3. If six were added to each value of the data in the table, the 15^{th} percentile of the new list of values is:

a. six
b. one
c. seven
d. eight

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is ten percent. However, in any year, the probability of water rationing is five percent.

4. What is the probability of both a drought and water rationing occurring?

a. 0.05
b. 0.01

c. 0.02
d. 0.30

5. Which of the following is true?

a. Drought and water rationing are independent events.
b. Drought and water rationing are mutually exclusive events.
c. None of the above

Use the following information to answer the next two exercises: Suppose that a survey yielded the following data:

gender	**apple**	**pumpkin**	**pecan**
female	40	10	30
male	20	30	10

Favorite Pie

6. Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is _____.

a. $\frac{40}{60}$
b. $\frac{60}{140}$
c. $\frac{120}{140}$
d. $\frac{100}{140}$

7. Suppose H_0 is: Favorite pie and gender are independent. The *p*-value is ______.

a. ≈ 0
b. 1
c. 0.05
d. cannot be determined

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number of people who watch the news at least once per week.

8. Which of the following statements is FALSE?

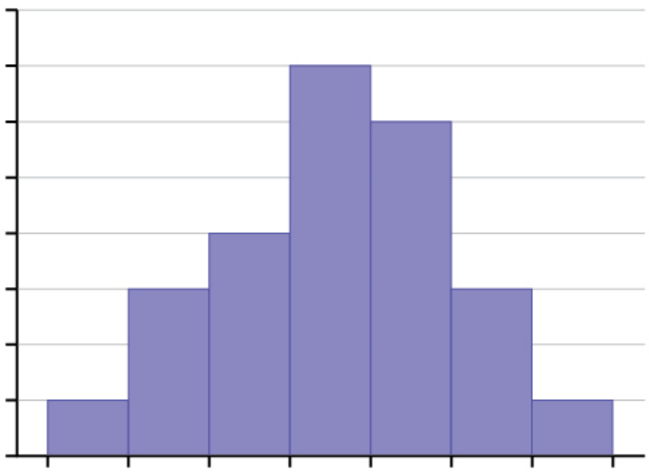
a. $X \sim B(14\ 0.60)$
b. The values for *x* are: {1 ,2 ,3 ,... ,14}.
c. $\mu = 8.4$
d. $P(X = 5) = 0.0408$

9. Find the probability that at least six adults watch the news at least once per week.

a. $\frac{6}{14}$
b. 0.8499
c. 0.9417

d. 0.6429

10. The following histogram is most likely to be a result of sampling from which distribution?



a. chi-square with *df* = 6
b. exponential
c. uniform
d. binomial

11. The ages of campus day and evening students is known to be normally distributed. A sample of six campus day and evening students reported their ages (in years) as: {18, 35, 27, 45, 20, 20}. What is the error bound for the 90% confidence interval of the true average age?

a. 11.2
b. 22.3
c. 17.5
d. 8.7

12. If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5% of the population values lie above:

a. –1.96.
b. 1.96.
c. 1.
d. –1.

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

13. What is the probability that one customer spends less than $72 in one trip to the supermarket?

a. 0.6321
b. 0.5000
c. 0.3714
d. 1

14. How much money altogether would you expect the next five customers to spend in one trip to the supermarket (in dollars)?

a. 72
b. $\frac{72^2}{5}$
c. 5184
d. 360

15. If you want to find the probability that the mean amount of money 50 customers spend in one trip to the supermarket is less than $60, the distribution to use is:

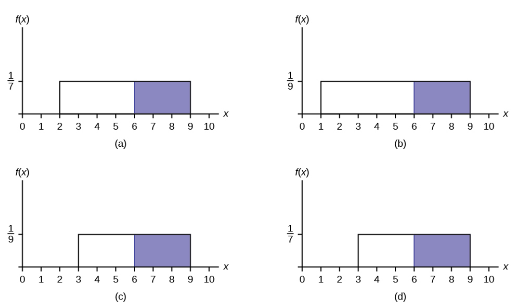
a. $N(72, 72)$
b. $N\left(72, \frac{72}{\sqrt{50}}\right)$
c. $Exp(72)$
d. $Exp\left(\frac{1}{72}\right)$

Use the following information to answer the next three exercises: The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from one to ten minutes.

16. What is the probability that a randomly chosen fourth grader takes more than seven minutes to take out the trash?

a. $\frac{3}{9}$
b. $\frac{7}{9}$
c. $\frac{3}{10}$
d. $\frac{7}{10}$

17. Which graph best shows the probability that a randomly chosen fourth grader takes more than six minutes to take out the trash given that he or she has already taken more than three minutes?



18. We should expect a fourth grader to take how many minutes to take out the trash?

a. 4.5
b. 5.5
c. 5
d. 10

Use the following information to answer the next three exercises: At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of five minutes and a standard deviation of 1.5 minutes.

19. What is the 90th percentile of waiting times (in minutes)?

a. 1.28

b. 90
c. 7.47
d. 6.92

20. The median waiting time (in minutes) for one student is:

a. 5.
b. 50.
c. 2.5.
d. 1.5.

21. Find the probability that the average wait time for ten students is at most 5.5 minutes.

a. 0.6301
b. 0.8541
c. 0.3694
d. 0.1459

22. A sample of 80 software engineers in Silicon Valley is taken and it is found that 20% of them earn approximately $50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn $50,000 per year is:

a. 16.
b. 0.2.
c. 1.
d. 0.95.

23. If $P(Z < z_{\alpha}) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to:

a. –1.
b. 0.1587.
c. 0.8413.
d. 1.

24. A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of:

a. independent groups.
b. two proportions.
c. matched pairs, dependent groups.
d. exclusive groups.

A math exam was given to all the third grade children attending ABC School. Two random samples of scores were taken.

	n	x	***s***
Boys	55	82	5
Girls	60	86	7

25. Which of the following correctly describes the results of a hypothesis test of the claim, “There is a difference between the mean scores obtained by third grade girls and boys at the 5% level of significance”?

a. Do not reject H_0. There is insufficient evidence to conclude that there is a difference in the mean scores.
b. Do not reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.
c. Reject H_0. There is insufficient evidence to conclude that there is no difference in the mean scores.
d. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is that:

a. there is insufficient information to conclude that the proportion for males is the same as the proportion for females.
b. there is insufficient information to conclude that the proportion for males is not the same as the proportion for females.
c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.
d. not enough information to make a conclusion.

27. From past experience, a statistics teacher has found that the average score on a midterm is 81 with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher’s claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

a. Yes
b. No
c. Not enough information given to solve the problem

28. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages with a standard deviation of two minutes. Machine II took an average of 28 minutes to load packages with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages with a standard deviation of one minute. Find the *p*-value when testing that the average loading times are the same.

a. *p*-value is close to zero
b. *p*-value is close to one
c. not enough information given to solve the problem

Use the following information to answer the next three exercises: A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees *x*	650	730	810	900	102	107	1150
Number of bathrooms *y*	40	50	54	61	82	110	121

29. Is the correlation between the number of employees and the number of bathrooms significant?

a. Yes
b. No
c. Not enough information to answer question

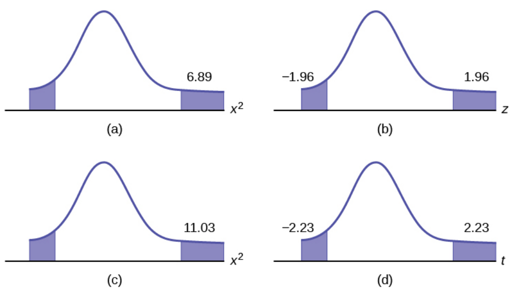
30. The linear regression equation is:

a. $\hat{y} = 0.0094 - 79.96x$
b. $\hat{y} = 79.96 + 0.0094x$
c. $\hat{y} = 79.96 - 0.0094x$
d. $\hat{y} = -0.0094 + 79.96x$

31. If a site has 1,150 employees, approximately how many bathrooms should it have?

a. 69
b. 91
c. 91,954
d. We should not be estimating here.

32. Suppose that a sample of size ten was collected, with $x = 4.4$ and $s = 1.4$. H_0: $\sigma^2 = 1.6$ vs. H_a: $\sigma^2 \neq 1.6$. Which graph best describes the results of the test?



Sixty-four backpackers were asked the number of days since their latest backpacking trip. The number of days is given in [link]:

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

33. Conduct an appropriate test to determine if the distribution is uniform.

a. The p-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.
b. The p-value is < 0.01. There is sufficient information to conclude the distribution is not uniform.
c. The p-value is between 0.01 and 0.10, but without alpha (α) there is not enough information
d. There is no such test that can be conducted.

34. Which of the following statements is true when using one-way ANOVA?

a. The populations from which the samples are selected have different distributions.
b. The sample sizes are large.
c. The test is to determine if the different groups have the same means.
d. There is a correlation between the factors of the experiment.

Practice Final Exam 2 Solutions

Solutions

1. b. parameter.

2. a.

3. c. seven

4. c. 0.02

5. c. none of the above

6. d. $\frac{100}{140}$

7. a. ≈ 0

8. b. The values for *x* are: {1, 2, 3,..., 14}

9. c. 0.9417.

10. d. binomial

11. d. 8.7

12. a. –1.96

13. a. 0.6321

14. d. 360

15. b. $N\left(72, \frac{72}{\sqrt{50}}\right)$

16. a. $\frac{3}{9}$

17. d.

18. b. 5.5

19. d. 6.92

20. a. 5

21. b. 0.8541

22. b. 0.2

23. a. –1.

24. c. matched pairs, dependent groups.

25. d. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

27. b. no

28. b. *p*-value is close to 1.

29. b. No

30. c. $\hat{y} = 79.96x - 0.0094$

31. d. We should not be estimating here.

32. a.

33. a. The p-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.

34. c. The test is to determine if the different groups have the same means.

Data Sets

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Race Lap Times (in seconds)

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44

$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00
$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	
$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

Group and Partner Projects

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.
_____ Decide what data you are going to study.

> **Note:**
> Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

_____ Are your data discrete or continuous? How do you know?
_____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?
_____ Conduct your survey. **Your data size must be at least 30.**
_____ Summarize your data in a chart with columns showing **data value,**

frequency, relative frequency and cumulative relative frequency.
Answer the following (rounded to two decimal places):

a. x = _____
b. s = _____
c. First quartile = _____
d. Median = _____
e. 70^{th} percentile = _____

____ What value is two standard deviations above the mean?

____ What value is 1.5 standard deviations below the mean?
____ Construct a histogram displaying your data.
____ In complete sentences, describe the shape of your graph.
____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.
____ Construct a box plot displaying your data.
____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- **____Cover sheet**: name, class time, and name of your study
- **____Summary page**: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
- **____URL** for data, if your data are from the World Wide Web

- ____**Chart of data, frequency, relative frequency, and cumulative relative frequency**
- ____**Page(s) of graphs:** histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance telephone call lasts.)
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.
____ Define (in words) the random variable for your data. X = ________
____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.

_____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. $\bar{x}$ = _______
b. s = _______

_____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

_____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.
_____ Uniform: $X \sim U$ ______________ Use the lowest and highest values as a and b.
_____ Normal: $X \sim N$ ______________ Use $\bar{x}$ to estimate for μ and s to estimate for σ.
_____ **Must** your data fit one of the above distributions? Explain why or why not.
_____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.
_____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = __________ (rounded to two decimal places) Note: $k = \bar{x} + (1.75)s$
_____ Determine the relative frequencies (RF) rounded to four decimal places.

Note:
Note
$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

a. $RF(X < k)$ = _______
b. $RF(X > k)$ = _______
c. $RF(X = k)$ = _______

Note:
Note
You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

_____ State the distribution: $X \sim$ __________
_____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
_____ Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k)$ = _______
b. $P(X > k)$ = _______
c. $P(X = k)$ = _______

_____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
_____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_______ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.
_______ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.

_______ List the 40 averages in order from smallest to largest.
_______ Define the random variable, X, in words. X = _________________
_______ State the approximate theoretical distribution of X. X ~ ________________
_______ Base this on the mean and standard deviation from your original data.
_______ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.
Calculate the value k (an X value) that is 1.75 standard deviations above the sample mean. k = _____ (rounded to two decimal places)
Determine the relative frequencies (*RF*) rounded to four decimal places.

a. $RF(X < k)$ = ________
b. $RF(X > k)$ = ________
c. $RF(X = k)$ = ________

Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k)$ = ________
b. $P(X > k)$ = ________
c. $P(X = k)$ = ________

_______ Draw the graph of the theoretical distribution of X.
_______ Compare the relative frequencies to the probabilities. Are the values close?
_______ Does it appear that the data of averages fit the distribution of X well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.
_______ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
_______ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would "it"

always happen? Why or why not?
______ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or X distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:
____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions
____ **URL** for data, if your data are from the World Wide Web
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.

____**Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.

____**Copy or print out the article** and include a copy in your project, along with the source.

____**State how you will collect your data.** (Convenience sampling is not acceptable.)

____**Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.

____**State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

____**Make two copies of the appropriate solution sheet.**

____**Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.

____**Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.

____**Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

a. Brief discussion of the article, including the source
b. Statement of the claim made in the article (one of the hypotheses).
c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified,

systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.

d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

____**Cover sheet** containing your name(s), class time, and the name of your study

____**Summary**, which includes all items listed on summary checklist

____**Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.

____**Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.

____**Raw data collected AND a table summarizing the sample data** (n, x and s; or x, n, and p', as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.

- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction
____State the bivariate data your group is going to study.

> **Note:**
> Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
____Conduct your survey. Your number of pairs must be at least 30.
____Print out a copy of your data.

Analysis
____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
____State the least squares line and the correlation coefficient.
____On your scatter plot, in a different color, construct the least squares line.
____Is the correlation coefficient significant? Explain and show how you determined this.
____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your *x*-values might be 1971, 1972, 1973, 1974, …, 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.
______Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
______Answer the following question, rounded to two decimal places:

a. Sample mean = _______
b. Sample standard deviation = _______
c. First quartile = _______
d. Third quartile = _______
e. Median = _______

f. 70th percentile = _______
g. Value that is 2 standard deviations above the mean = _______
h. Value that is 1.5 standard deviations below the mean = _______

______Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
______In complete sentences, describe the shape of your histogram.
______Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in Descriptive Statistics (since you are now using univariate data) to determine which values might be outliers.
______Construct a box plot of your data.
______Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.
______Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: ___________ (keep a copy for your records)
- Part I, Analysis: ___________ (keep a copy for your records)

- Entire Project, typed and stapled: ___________

_____ Cover sheet: names, class time, and name of your study

_____ Part I: label the sections “Intro” and “Analysis.”

____ Part II:

____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.

____ All graphs requested in the project

____ All calculations requested to support questions in data

____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

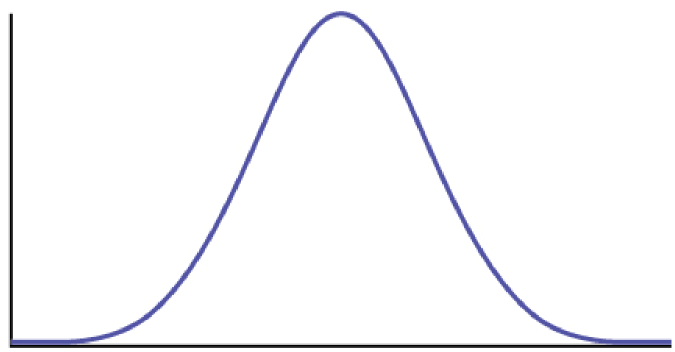
> **Note:**
> Note
> Include answers to ALL questions asked, even if not explicitly repeated in the items above.

Solution Sheets

Hypothesis Testing with One Sample

Class Time: ____________________________
Name: ____

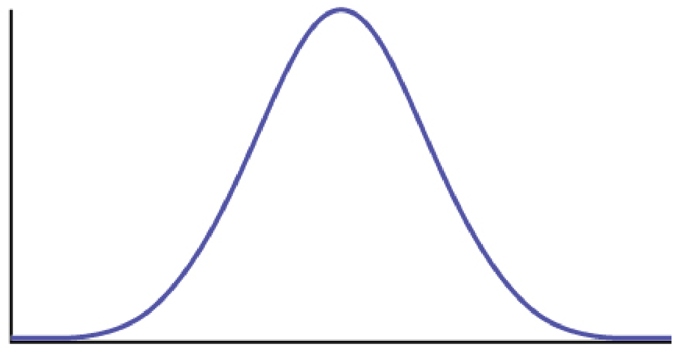
a. H_0: ________
b. H_a: ________
c. In words, **CLEARLY** state what your random variable X or P' represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one or two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

   i. Alpha: ________
   ii. Decision: ________
   iii. Reason for decision: ________
   iv. Conclusion: ________

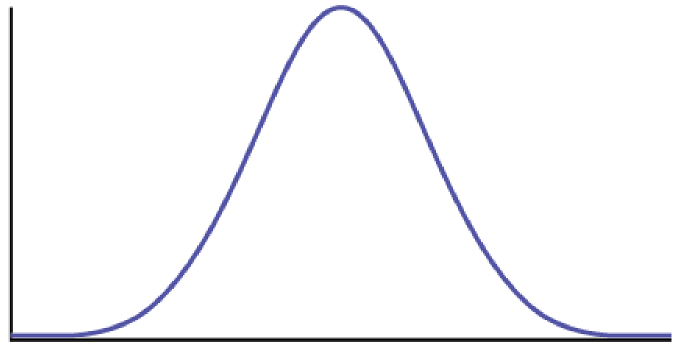
i. Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.



Hypothesis Testing with Two Samples

Class Time: ____________________________
Name: ____

a. H_0: ________
b. H_a: ________
c. In words, **clearly** state what your random variable $X_1 - X_2$, $P'_1 - P'_2$ or X_d represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the p-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

   a. Alpha: ________
   b. Decision: ________
   c. Reason for decision: ________
   d. Conclusion: ________

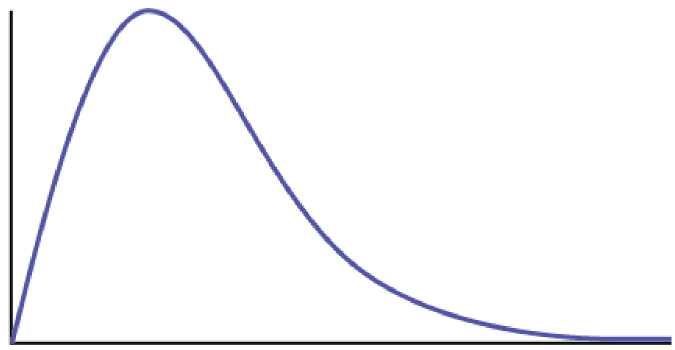
i. In complete sentences, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: ____________________________

Name: _______________________________________

a. H_0: ________
b. H_a: ________
c. What are the degrees of freedom?
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences.**

   i. Alpha: ________
   ii. Decision: ________
   iii. Reason for decision: ________
   iv. Conclusion: ________

F Distribution and One-Way ANOVA

Class Time: ____________________________
Name: _______________________________________

a. H_0: ________
b. H_a: ________
c. $df(n)$ = _______ $df(d)$ = ________
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value?
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.

a. Alpha: ________
b. Decision: ________
c. Reason for decision: ________
d. Conclusion: ________

Mathematical Phrases, Symbols, and Formulas

English Phrases Written Mathematically

When the English says:	**Interpret this as:**
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$

When the English says:	Interpret this as:
There are fewer *X* than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Formulas

Formula 1: Factorial

$n! = n(n-1)(n-2)\ldots(1)$

$0! = 1$

Formula 2: Combinations

$\binom{n}{r} = \frac{n!}{(n-r)!r!}$

Formula 3: Binomial Distribution

$X \sim B(n, p)$

$P(X = x) = \binom{n}{x} p^x q^{n-x}$, for $x = 0, 1, 2, \ldots, n$

Formula 4: Geometric Distribution

$X \sim G(p)$

$P(X = x) = q^{x-1}p$, for $x = 1, 2, 3, \ldots$

Formula 5: Hypergeometric Distribution

$X \sim H(r, b, n)$

$$P(X = x) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}}\right)$$

Formula 6: Poisson Distribution

$X \sim P(\mu)$

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

Formula 7: Uniform Distribution

$X \sim U(a, b)$

$f(X) = \frac{1}{b-a}, a < x < b$

Formula 8: Exponential Distribution

$X \sim Exp(m)$

$f(x) = me^{-mx}m > 0, x \geq 0$

Formula 9: Normal Distribution
$X \sim N(\mu, \sigma^2)$

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$

Formula 10: Gamma Function

$$\Gamma(z) = \int_{\infty}^{0} x^{z-1}e^{-x}dx \; z > 0$$

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$\Gamma(m+1) = m!$ for m, a nonnegative integer

otherwise: $\Gamma(a+1) = a\Gamma(a)$

Formula 11: Student's *t*-distribution

$X \sim t_{df}$

$$f(x) = \frac{\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\mathrm{n}\pi}\,\Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$Z \sim N(0,1\ ),\ Y \sim X^2_{df}$, n = degrees of freedom

Formula 12: Chi-Square Distribution

$X \sim X^2_{df}$

$f(x) = \dfrac{x^{\frac{n-2}{2}} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$, $x > 0$, n = positive integer and degrees of freedom

Formula 13: F Distribution

$X \sim F_{df(n),df(d)}$

$df(n) =$ degrees of freedom for the numerator

$df(d) =$ degrees of freedom for the denominator

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\left(\frac{u}{2}-1\right)} \left[1 + \left(\frac{u}{v}\right) x^{-0.5(u+v)}\right]$$

$X = \frac{Y_u}{W_v}$, Y, W are chi-square

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\ }$	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	IQR	interquartile range	$Q_3 – Q_1 = IQR$
Descriptive Statistics	$\overline{x}$	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Descriptive Statistics	**s** s_x *sx*	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s squared	sample variance
Descriptive Statistics	σ σ_x *σx*	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Probability Topics	$P(A\|B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	$P(A')$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	$X \sim$	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	$\neq$	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	*pdf*	prob. density function	same
Continuous Random Variables	U	uniform distribution	same
Continuous Random Variables	*Exp*	exponential distribution	same
Continuous Random Variables	k	k	critical value

Chapter (1st used)	Symbol	Spoken	Meaning
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	m	m	decay rate (for exp. dist.)
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z-score	same
The Normal Distribution	Z	standard normal dist.	same
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem	$\overline{X}$	X-bar	the random variable X-bar
The Central Limit Theorem	μ_x	mean of X	the average of X

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
The Central Limit Theorem	$\mu_{\overline{x}}$	mean of *X*-bar	the average of *X*-bar
The Central Limit Theorem	σ_x	standard deviation of *X*	same
The Central Limit Theorem	$\sigma_{\overline{x}}$	standard deviation of *X*-bar	same
The Central Limit Theorem	ΣX	sum of *X*	same
The Central Limit Theorem	Σx	sum of *x*	same
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same

Chapter (1st used)	Symbol	Spoken	Meaning
Confidence Intervals	t	Student's *t*-distribution	same
Confidence Intervals	df	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with *a*/2 area in right tail	same
Confidence Intervals	$p\prime$; $\widehat{p}$	*p*-prime; *p*-hat	sample proportion of success
Confidence Intervals	$q\prime$; $\hat{q}$	*q*-prime; *q*-hat	sample proportion of failure
Hypothesis Testing	H_0	*H*-naught, *H*-sub 0	null hypothesis
Hypothesis Testing	H_a	*H*-*a*, *H*-sub *a*	alternate hypothesis
Hypothesis Testing	H_1	*H*-1, *H*-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X1} - \overline{X2}$	*X*1-bar minus *X*2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	*mu*-1 minus *mu*-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	*P*1-prime minus *P*2-prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	*p*1 minus *p*2	difference in population proportions
Chi-Square Distribution	X^2	*Ky*-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Linear Regression and Correlation	$y = a + bx$	y equals a plus b-x	equation of a line
Linear Regression and Correlation	$\hat{y}$	y-hat	estimated value of y
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ε	error	same
Linear Regression and Correlation	SSE	Sum of Squared Errors	same
Linear Regression and Correlation	$1.9s$	1.9 times s	cut-off value for outliers

Chapter (1st used)	Symbol	Spoken	Meaning
F-Distribution and ANOVA	F	F-ratio	F-ratio

Symbols and their Meanings

Notes for the TI-83, 83+, 84, 84+ Calculators

Quick Tips

Legend

- 

  represents a button press
- [ ] represents yellow command or green letter behind a key
- < > represents items on the screen

To adjust the contrast
Press

, then hold

to increase the contrast or

to decrease the contrast.

To capitalize letters and words
Press

to get one capital letter, or press

, then

to set all button presses to capital letters. You can return to the top-level button values by pressing

ALPHA

again.

To correct a mistake
If you hit a wrong button, just hit

CLEAR

and start again.

To write in scientific notation
Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- 4.321 E 4 = 4.321×10^4
- 4.321 E –4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another:
Both calculators: Insert your respective end of the link cable cable and press

2nd

, then `[LINK]`.

Calculator receiving information:

Use the arrows to navigate to and select`<RECEIVE>`
Press ENTER .

Calculator sending information:

Press appropriate number or letter.
Use up and down arrows to access the appropriate item.

Press ENTER to select item to transfer.

Press right arrow to navigate to and select `<TRANSMIT>`.
Press ENTER.

> **Note:**
> Note
> ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press

2nd

, then `[QUIT]` to exit when done.

Manipulating One-Variable Statistics

> **Note:**
> Note
> These directions are for entering data with the built-in statistical program.

Data	**Frequency**

Data	Frequency
–2	10
–1	3
0	4
1	5
3	8

Sample DataWe are manipulating one-variable statistics.

To begin:

1. Turn on the calculator.

   ON

2. Access statistics mode.

   STAT

3. Select <4:ClrList> to clear data from lists, if desired.

   4

   ,

   ENTER

4. Enter list [L1] to be cleared.

   2nd

   , [L1] ,

   ENTER

5. Display last instruction.

   2nd

   , [ENTRY]

6. Continue clearing remaining lists in the same fashion, if desired.

   ◀

   ,

   2nd

   , [L2] ,

   ENTER

7. Access statistics mode.

   STAT

8. Select <1:Edit . . .>

   ENTER

9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1]).

   - Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad).

     (−)

     ,

     9

     ,

     ENTER

- Continue in the same manner until all data values are entered.

10. In [L2], enter the frequencies for each data value in [L1].

    - Type in a frequency and enter it. (If a data value appears only once, the frequency is "1").

      4

      ,

      ENTER

    - Continue in the same manner until all data values are entered.

11. Access statistics mode.

    STAT

12. Navigate to <CALC>.

13. Access <1:1-var Stats>.

    ENTER

14. Indicate that the data is in [L1]...

    2nd

    , [L1] ,

    ,

15. ...and indicate that the frequencies are in [L2].

    2nd

    , [L2] ,

    ENTER

16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

> **Note:**
> Note
> We will assume that the data is already entered.

We will construct two histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

   , `[STAT PLOT]`
2. Select `<1:plot 1>` to access plotting - first graph.

3. Use the arrows navigate go to `<ON>` to turn on Plot 1.
   `<ON>` ,

4. Use the arrows to go to the histogram picture and select the histogram.

5. Use the arrows to navigate to `<Xlist>`.
6. If "L1" is not selected, select it.

2nd

, [L1] ,

ENTER

7. Use the arrows to navigate to <Freq>.

8. Assign the frequencies to [L2].

   2nd

   , [L2] ,

   ENTER

9. Go back to access other graphs.

   2nd

   , [STAT PLOT]
10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.

   Y=

2. Select each equal sign (=).

   ▼

   ▶

   ENTER

3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.

2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.

2. Select `<9:ZoomStat>`.

3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access window mode to set the graph parameters.

2.
   - $X_{\min} = -2.5$
   - $X_{\max} = 3.5$
   - $X_{scl} = 1$ (width of bars)
   - $Y_{\min} = 0$
   - $Y_{\max} = 10$
   - $Y_{scl} = 1$ (spacing of tick marks on *y*-axis)
   - $X_{res} = 1$

3. Access graphing mode to see the histogram.

   GRAPH

To draw box plots:

1. Access graphing mode.

   2nd

   , [STAT PLOT]

2. Select <1:Plot 1> to access the first graph.

   ENTER

3. Use the arrows to select <ON> and turn on Plot 1.

   ENTER

4. Use the arrows to select the box plot picture and enable it.

   ENTER

5. Use the arrows to navigate to <Xlist>.

6. If "L1" is not selected, select it.

   2nd

   , [L1] ,

   ENTER

7. Use the arrows to navigate to <Freq>.

8. Indicate that the frequencies are in [L2].

   2nd

   , [L2] ,

ENTER

9. Go back to access other graphs.

   2nd

   , `[STAT PLOT]`
10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.

11. View the box plot.

   GRAPH

   , `[STAT PLOT]`

Linear Regression

Sample Data

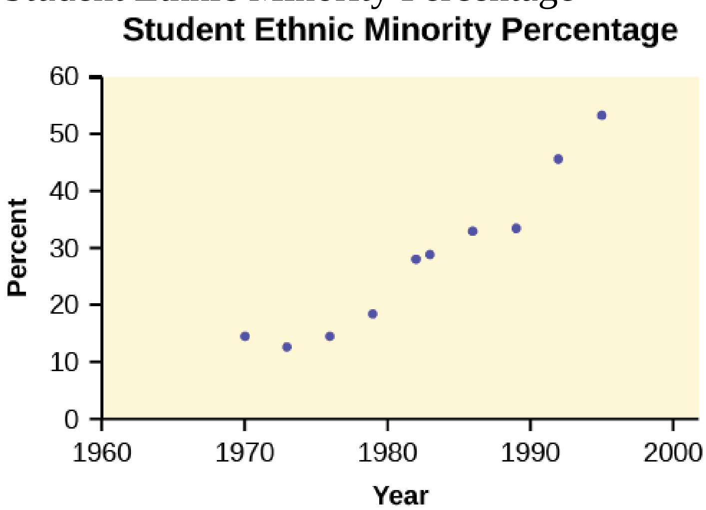
The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 was:

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16

Year	Student Ethnic Minority Percentage
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

Student Ethnic Minority Percentage



By hand, verify the scatterplot above.

Note:

Note
The TI-83 has a built-in linear regression feature, which allows the data to be edited.The *x*-values will be in `[L1]`; the *y*-values in `[L2]`.

To enter data and do linear regression:

1. ON Turns calculator on.

   ON

2. Before accessing this program, be sure to turn off all plots.

   - Access graphing mode.

     2nd

     , `[STAT PLOT]`

   - Turn off all plots.

     4

     ,

     ENTER

3. Round to three decimal places. To do so:

   - Access the mode menu.

     MODE

     , `[STAT PLOT]`

   - Navigate to `<Float>` and then to the right to `<3>`.

     ▼

     ▶

- All numbers will be rounded to three decimal places until changed.

  ENTER

4. Enter statistics mode and clear lists `[L1]` and `[L2]`, as describe previously.

   STAT

   ,

   4

5. Enter editing mode to insert values for *x* and *y*.

   STAT

   ,

   ENTER

6. Enter each value. Press

   ENTER

   to continue.

To display the correlation coefficient:

1. Access the catalog.

   2nd

   , `[CATALOG]`

2. Arrow down and select `<DiagnosticOn>`

   ▼

   ... ,

ENTER

,

ENTER

3. r and r^2 will be displayed during regression calculations.

4. Access linear regression.

STAT

▶

5. Select the form of *y* = *a* + *bx*.

8

,

ENTER

The display will show:

LinReg

- *y* = *a* + *bx*
- *a* = –3176.909
- *b* = 1.617
- *r* = 2 0.924
- *r* = 0.961

This means the Line of Best Fit (Least Squares Line) is:

- *y* = –3176.909 + 1.617*x*
- Percent = –3176.909 + 1.617 (year #)

The correlation coefficient *r* = 0.961

To see the scatter plot:

1. Access graphing mode.

   2nd

   , [STAT PLOT]

2. Select <1:plot 1> To access plotting - first graph.

   ENTER

3. Navigate and select <ON> to turn on Plot 1.
   <ON>

   ENTER

4. Navigate to the first picture.

5. Select the scatter plot.

   ENTER

6. Navigate to <Xlist>.
7. If [L1] is not selected, press

   2nd

   , [L1] to select it.

8. Confirm that the data values are in [L1].
   <ON>

   ENTER

9. Navigate to <Ylist>.

10. Select that the frequencies are in [L2].

    2nd

, `[L2]` ,

ENTER

11. Go back to access other graphs.

    2nd

    , `[STAT PLOT]`
12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.

    WINDOW

    - $X_{\min} = 1970$
    - $X_{\max} = 2000$
    - $X_{scl} = 10$ (spacing of tick marks on *x*-axis)
    - $Y_{\min} = -0.05$
    - $Y_{\max} = 60$
    - $Y_{scl} = 10$ (spacing of tick marks on *y*-axis)
    - $X_{res} = 1$

14. Be sure to deselect or clear all equations before graphing, using the instructions above.
15. Press the graph button to see the scatter plot.

    GRAPH

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.

   Y=

2. Access the vars menu and navigate to `<5: Statistics>`.

   VARS

   ,

3. Navigate to `<EQ>`.

4. `<1: RegEQ>` contains the regression equation which will be entered in Y1.

5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.

To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

, `[LIST]`, `<RESID>`

2. Confirm twice to view the list of residuals. Use the arrows to select them.

,

3. The critical point for an outlier is: $1.9V\frac{\text{SSE}}{n-2}$ where:

    - n = number of pairs of data
    - SSE = sum of the squared errors
    - $\sum$ residual^2

4. Store the residuals in `[L3]`.

,

2nd

, [L3] ,

ENTER

5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$

2nd

, [L3] ,

x^2

,

÷

,

8

6. Store this value in [L4].

STO►

,

2nd

, [L4] ,

ENTER

7. Calculate the critical value using the equation above.

1

,

.

,

9

,

X

,

2nd

, `[V]` ,

2nd

, `[LIST]`

▶

,

▶

,

5

,

2nd

, `[L4]` ,

)

,

)

,

ENTER

8. Verify that the calculator displays: 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in `[L3]` to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of *y* for various *x*-values:
There are various ways to determine estimates for "*y*." One way is to substitute values for "*x*" in the equation. Another way is to use the

TRACE

on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access `DISTR` (for "Distributions").

For technical assistance, visit the Texas Instruments website at http://www.ti.com and enter your calculator model into the "search" box.

Binomial Distribution

- `binompdf(n,p,x)` corresponds to $P(X = x)$
- `binomcdf(n,p,x)` corresponds to $P(X \leq x)$
- To see a list of all probabilities for *x*: 0, 1, . . . , *n*, leave off the "`x`" parameter.

Poisson Distribution

- `poissonpdf(λ,x)` corresponds to $P(X = x)$
- `poissoncdf(λ,x)` corresponds to $P(X \leq x)$

Continuous Distributions (general)

- $-\infty$ uses the value –1EE99 for left bound
- ∞ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x,μ,σ)` yields a probability density function value (only useful to plot the normal curve, in which case "`x`" is the variable)
- `normalcdf(left bound, right bound, μ, σ)` corresponds to *P*(left bound < *X* < right bound)
- `normalcdf(left bound, right bound)` corresponds to *P*(left bound < *Z* < right bound) – standard normal
- `invNorm(p,μ,σ)` yields the critical value, *k*: *P*(*X* < *k*) = *p*
- `invNorm(p)` yields the critical value, *k*: *P*(*Z* < *k*) = *p* for the standard normal

Student's *t*-Distribution

- `tpdf(x,df)` yields the probability density function value (only useful to plot the student-*t* curve, in which case "`x`" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to *P*(left bound < *t* < right bound)

Chi-square Distribution

- `X²pdf(x,df)` yields the probability density function value (only useful to plot the chi^2 curve, in which case "`x`" is the variable)
- `X²cdf(left bound, right bound, df)` corresponds to *P*(left bound < X^2 < right bound)

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value (only useful to plot the *F* curve, in which case "`x`" is the variable)
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to *P*(left bound < *F* < right bound)

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press `DATA` to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing `STAT` once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known.
- `TInterval` is the confidence interval for mean when σ is unknown; *s* estimates σ.
- `1-PropZInt` is the confidence interval for proportion.

Note:
Note
The confidence levels should be given as percents (ex. enter "`95`" or "`.95`" for a 95% confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known.
- `T-Test` is the hypothesis test for single mean when σ is unknown; *s* estimates σ.
- `2-SampZTest` is the hypothesis test for two independent means when both σ's are known.
- `2-SampTTest` is the hypothesis test for two independent means when both σ's are unknown.
- `1-PropZTest` is the hypothesis test for single proportion.
- `2-PropZTest` is the hypothesis test for two proportions.
- X^2`-Test` is the hypothesis test for independence.

- `X²GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

Note:
Note
Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "`μ∅`" represents the null hypothesis. For a test of a single proportion, "`p∅`" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

Tables

The module contains links to government site tables used in statistics.

Note:
Note
When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- Student *t* table
- Normal table
- Chi-Square table
- F-table
- All four tables can be accessed by going to

95% Critical Values of the Sample Correlation Coefficient Table

- 95% Critical Values of the Sample Correlation Coefficient